The genome and epigenome of the European ash tree (*Fraxinus excelsior*)

Elizabeth Sollars

Thesis submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy

School of Biological and Chemical Sciences
Queen Mary University of London

Supervisor: Dr. Richard Buggs

April 2017

Statement of Originality

I, Elizabeth Sollars, confirm that the research included within this thesis is my own work or that where it has been carried out in collaboration with, or supported by others, that this is duly acknowledged below and my contribution indicated. Previously published material is also acknowledged below.

I attest that I have exercised reasonable care to ensure that the work is original, and does not to the best of my knowledge break any UK law, infringe any third partys copyright or other Intellectual Property Right, or contain any confidential material.

I accept that the College has the right to use plagiarism detection software to check the electronic version of the thesis.

I confirm that this thesis has not been previously submitted for the award of a degree by this or any other university.

Signature: E. Sollars
Date: 26th April 2017

Abstract

European ash trees (*Fraxinus excelsior*) are under threat from the fungal pathogen *Hymenoscyphus fraxineus* causing ash dieback disease (ADB). Previous research has shown heritable variation in ADB susceptibility in natural ash populations. Prior to this project, very little genetic data were available for ash, thus hampering efforts to identify markers associated with susceptibility. In this thesis, I have presented nuclear and organellar assemblies of the 880 Mbp *F. excelsior* genome, with a combined N50 scaffold size of over 100 kbp. Using Ks distributions for six plant species, I found evidence for two whole genome duplication (WGD) events in the history of the ash lineage, one potentially shared with olive (Ks ∼0.4), and one potentially with other members of the Lamiales order (Ks ∼0.7). Using a further 38 genome sequences from trees originating throughout Europe, I found little evidence of any population structure throughout the European range of *F.excelsior*, but find a substantial decrease in effective population size, both in the distant (from ∼10 mya) and recent past. Linkage disequilibrium is low at small distances between loci, with an r^2 of 0.15 at a few hundred bp, but decays slowly from this point. From whole genome DNA methylation data of twenty *F. excelsior* and *F. mandshurica* trees, I identified 665 Differentially Methylated Regions (DMRs) between those with high and low ADB susceptibility. Of genes putatively duplicated in historical WGD events, an average of 25.9% were differentially methylated in at least one cytosine context, possibly indicative of unequal silencing. Finally, I found some variability in methylation patterns among clonal replicates (Pearson's correlation coefficient ∼0.960), but this was less than the variability found between different genotypes (∼0.955). The results from this project and the genome sequence especially, will be valuable to researchers aiming to breed or select ash trees with low susceptibility to ADB.

Acknowledgments

I have been exceptionally proud to study my PhD as a member of the EU-funded ITN network, INTERCROSSING. In no other forum would I have had the chance to meet such a diverse and intelligent group of people and visit such beautiful European cities. Therefore, I would like to thank all other twelve members of INTERCROSSING, and their PIs, for enhancing my knowledge in a wide range of topics and enjoying many workshops together. My special thanks go to my INTERCROSSING partner, Jasmin Zohren, for being a wonderful colleague and friend, and to Richard Nichols, for organising and co-ordinating the network. I would like to thank my supervisor Richard Buggs for his encouragement and support throughout my PhD. In addition, I have received advice and help from members of the Buggs lab numerous times, so I thank all of them, especially Laura Kelly and Endymion Cooper.

I would also like to thank all the employees at CLC bio in Aarhus, who went to much effort to make a newcomer feel at home, and made every day at work a fun one. To have had the chance to live and work in the city of Aarhus has been truly life-changing. I would also like to thank the members of the Aarhus 1900 Triathlon and Run for Friendship clubs, for being great friends and making training so fun.

As always, thanks go to my family and friends for their continued love and support.

Funding

This work was supported by the EU FP7-PEOPLE project 'INTERCROSSING', ID:289974. Sequencing of the reference tree was funded by NERC emergency grant NE/K01112X/1.

Contents

1 Introduction to ash trees and the threat from ash dieback disease 13

1.1 Biology and geography of *Fraxinus excelsior* 13
1.2 Value of ash 15
1.3 *Hymenoscyphus fraxineus*, the causative agent of ash dieback 15
1.4 Natural genetic variation in susceptibility to ADB 18
1.5 Current status of ADB research and future directions 20

2 Introduction to current genome projects of forest trees 23

2.1 Challenges in sequencing and assembling plant genomes 24
2.1.1 Heterozygosity 24
2.1.2 Large genome size 25
2.2 Current forest tree genome projects 25
2.2.1 The first tree genome; *Populus trichocarpa* 26
2.2.2 Large and complex genomes - Gymnosperms *Pinus taeda* and *Picea* spp. 28
2.2.3 Breeding for desirable traits - *Salix* spp. 29
2.2.4 Disease resistance - *Castanea mollissima* 30
2.2.5 Population biology and phylogeography - *Betula* spp. 30
2.2.6 Tree genome databases 31
2.3 Conclusions 32

3 *De novo* genome assembly and annotation of a British *Fraxinus excelsior* tree 34

3.1 Introduction to genome assembly and finishing methods 34
3.1.1 de Bruijn graphs vs Overlap Layout Consensus methods 34
3.1.2 Scaffolding and gap filling 37
3.1.3 Assembly verification and comparison 38
3.2 Methods 39
3.2.1 DNA Extraction and sequencing 39
3.2.2 *De novo* assembly 40
3.2.3 Gene annotation 42
3.3 *De novo* assembly results 42
3.3.1 Overall comparison of released assemblies 42
3.3.2 Testing different software 43
3.4 RNA-seq aided annotation of genes 45
3.5 Conclusion 46

4 Assembly and annotation of organellar genomes from whole genome sequencing reads **49**
4.1 Methods . . . 50
4.1.1 Generating k-mer distributions . . . 50
4.1.2 Extracting organellar reads . . . 50
4.1.3 Plastid genome assembly and annotation . . . 51
4.1.4 Mitochondrial genome assembly and annotation . . . 52
4.2 Results . . . 53
4.2.1 K-mer distributions reveal peaks of organellar sequence coverage . . . 53
4.2.2 Assembly and annotation of the plastid genome . . . 54
4.2.3 Assembly of the mitochondrial genome using a map, extend and join method . . . 56
4.3 Conclusion . . . 58

5 Analysis of whole genome duplications in *Fraxinus excelsior* **60**
5.1 Introduction . . . 60
5.1.1 A rich history of whole genome duplications (WGD) in plants . . . 60
5.1.2 Overview of the Ks method . . . 63
5.1.3 Correcting for redundant Ks values in homeolog groups . . . 64
5.2 Methods . . . 64
5.3 Evidence for two WGD events in the ash lineage . . . 66
5.4 Conclusion . . . 68

6 Population structure among European ash trees **72**
6.1 Introduction . . . 72
6.1.1 Current population research on ash . . . 72
6.1.2 Approaches for analyzing population structure . . . 75
6.1.3 Approaches for estimating past N_e . . . 75
6.2 Methods . . . 77
6.2.1 Locations and origins of samples . . . 77
6.2.2 DNA sequencing and variant calling methods . . . 78
6.2.3 Population structure methods . . . 79
6.2.4 Effective population size methods . . . 80
6.3 Results and Discussion . . . 81
6.3.1 Analysis of population structure . . . 81
6.3.2 Estimating effective population size history . . . 85
6.4 Conclusion and future directions . . . 89

7 Epigenetic variation in isogenic samples **93**
7.1 Introduction . . . 93
7.1.1 DNA methylation in plants . . . 93
7.1.2 Background to the ash methylome project . . . 94
7.1.3 Bisulphite sequencing and alignment software . . . 95
7.2 Methods . . . 97
7.2.1 Description of samples and genotypes . . . 97
7.2.2 DNA extraction, bisulphite conversion and sequencing . . . 98

7.2.3 Data QC and read mapping . . . 98
7.2.4 Data analysis methods . . . 99
7.3 Results and Discussion . . . 101
7.3.1 Landscape of DNA methylation across the ash genome . . . 101
7.3.2 Many homeologs retained after WGD are differentially methylated . . 105
7.3.3 Methylation differences between two *Fraxinus* species and within genotypes . . . 110
7.3.4 Methylation in genes relating to ADB susceptibility . . . 115
7.4 Conclusion . . . 117

8 Conclusions and further research **121**
8.1 Summary of results . . . 121
8.2 Future research using the ash genome . . . 122

Bibliography **126**

Appendix **146**
Appendix 1: Published article on the ash tree genome . . . 147
Appendix 2: Book chapter on genomics projects of angiosperm trees . . . 170

List of Figures

1.1 Phylogeny of *Fraxinus* genus . . . 14
1.2 European distribution of *Fraxinus excelsior* . . . 15
1.3 Distribution of ash in Great Britain . . . 16
1.4 Life cycle of *Hymenoscyphus fraxineus* . . . 17
1.5 Year of first identification of ash dieback in European countries . . . 18
1.6 Distribution of confirmed ADB sites across UK . . . 19
1.7 *Fraxinus* phylogeny with ADB susceptibility . . . 21

3.1 De Bruijn graph assembly process . . . 35
3.2 FR curve of software comparison for BATG-0.3 . . . 45
3.3 FR curve for software comparisons for BATG-0.5 . . . 46

4.1 The map-extend-join method . . . 53
4.2 Mapped reads used to check a join between contigs . . . 53
4.3 Peak of k-mer frequency at 700x shows coverage of mitochondrial genome . . . 54
4.4 Peak of k-mer frequency at 4200x shows coverage of plastid genome . . . 54
4.5 Alignment of assembled plastid contigs against olive plastid sequence . . . 55
4.6 Gene annotation and structure of the chloroplast chromosome . . . 56
4.8 Alignment of *atp6* gene from six plant species . . . 57
4.7 Genes annotated on mitochondrial chromosome assembly . . . 58

5.1 Occurrences of WGD events in the plant kingdom . . . 61
5.2 Ks distribution of *Arabidopsis thaliana* . . . 63
5.3 WGD event history in core angiosperms . . . 65
5.4 Ks plots for seven plant species . . . 70
5.5 Ks plot from tomato genome paper . . . 71
5.6 Updated WGD phylogeny of seven species . . . 71

6.1 Geographical distribution of ash chloroplast haplotypes . . . 74
6.2 European distribution of *F. excelsior* and *F. angustifolia* . . . 74
6.3 PSMC method . . . 76
6.4 Original locations of 38 ash trees . . . 77
6.5 Delta K plot of STRUCTURE results . . . 82
6.6 STRUCTURE results from three independent SNP sets . . . 82
6.7 Map of STRUCTURE results . . . 83
6.8 Median-joining network of plastid haplotype sequence alignment . . . 84

6.9 Map of plastid haplotype groups . . . 84
6.10 PC1 vs PC2 clusters trees similarly to STRUCTURE . . . 85
6.11 PC2 vs PC3 clusters trees similarly to plastid haplotype network . . . 85
6.12 N_e change between 10 mya and 200 kya . . . 86
6.13 N_e change in the recent past . . . 87
6.14 Linkage disequilibrium decay over increasing distance between loci . . . 88
6.15 Additional comparison of forest tree LD values . . . 89
6.16 LD decay in three *Populus* species . . . 90

7.1 Process of bisulphite conversion . . . 96
7.2 Percentage of methylated cytosines in each sequence context . . . 103
7.3 Proportion of methylated bases in each context occurring at various methylation levels . . . 103
7.4 Methylation levels across genomic regions . . . 104
7.5 Methylation across genes, TEs, and flanking regions . . . 104
7.6 Differential methylation in homeologs of one *F. excelsior* sample . . . 106
7.7 Power curves for detecting methylation differences between homeologs . . . 108
7.8 Principal Components Analysis of methylation values for all samples . . . 111
7.9 Principal Components Analysis of methylation values for high coverage *F. excelsior* samples . . . 111
7.10 Hierarchical cluster of high coverage samples using methylation values . . . 113
7.11 Correlation matrix heatmap of all samples using methylation values . . . 114
7.12 Two significantly differentially methylated genes known to be associated with ADB susceptibility . . . 115
7.13 Power curves for DMR test . . . 118

List of Tables

2.1 Forest tree species with genome sequences available 26
2.2 Angiosperm forest tree species with genome-wide data available 27

3.1 Methods for five assembly versions 41
3.2 Sequencing yield of 2451S 43
3.3 Comparison of five genome assemblies 43
3.4 Statistics of assemblies for testing version 0.3 44
3.6 RNA sequencing yield of five ash samples 46
3.5 Statistics of assemblies for testing version 0.5 (at end of chapter) 48

4.1 Statistics of mitochondrial genome assembly 57

6.1 Source locations, sequencing and mapping results of 38 trees used in population analyses 78

7.1 Description of twenty ash trees using WGBS study 98
7.2 WGBS yield and mapping results for twenty ash samples 102
7.3 Twenty genes with highest density of N-DMPs 105
7.4 Percentage of significantly differentially methylated homeolog pairs for each *F. excelsior* tree 106
7.5 Most significant differentially methylated homeologs 109
7.6 Genomic positions with most effect on separation of two *Fraxinus* species . . 112
7.7 Methylation levels in twenty genes known to be associated with ADB susceptibility 116
7.8 Genes associated with most significant DMRs 120

List of abbreviations

ADB	Ash Dieback (disease)
AFS	Allele Frequency Spectrum
BAC	Bacterial Artificial Chromosome
BATG	British Ash Tree Genome
BLAST	Basic Local Alignment Search Tool
CDS	Coding DNA Sequence
cpDNA	chloroplast DNA
CTAB	Cetyl trimethylammonium bromide
DMP	Differentially Methylated Position
DMR	Differentially Methylated Region
EAB	Emerald Ash Borer
EST	Expressed Sequence Tag
FAD	Flavin Adenine Dinucleotide
FDR	False Discovery Rate
FR	Feature Response
GBS	Genotyping by Sequencing
GDS	Genomic Data Structure
GEM	Gene Expression Marker
GFF	General Feature Format
GO	Gene Ontology (term): (BP: Biological Process, CC: Cellular Component, MF: Molecular Function)
HMM	Hidden Markov Model
KOGs	euKaryotic Cluster of Orthologous Groups
Ks	Synonymous substitutions per synonymous site
LD	Linkage Disequilibrium
LGM	Last Glacial Maximum
LJD	Long Jumping Distance
LSCR	Large Single Copy Region
MAF	Minimum Allele Frequency
MNV	Multiple Nucleotide Variant
mtDNA	mitochondrial DNA
N-DMP	Non-Differentially Methylated Position
N_e	Effective population size
NGS	Next Generation Sequencing
OLC	Overlap Layout Consensus
PCA	Principal Components Analysis
PC	Principal Component
PCR	Polymerase Chain Reaction
PSMC	Pairwise Sequentially Markovian Coalescent (model)
QC	Quality Control
QTL	Quantitative Trait Loci
RAD	Restriction Site Associated DNA
RdDM	RNA-directed DNA Methylation
RFLP	Restriction Fragment Length Polymorphism
RRBS	Reduced Representation Bisulphite Sequencing
SFS	Site Frequency Spectrum
SNP	Single Nucleotide Polymorphism
SNV	Single Nucleotide Variant
SSR	Simple Sequence Repeat
TE	Transposable Element
TMRCA	Time to Most Recent Common Ancestor
VCF	Variant Call Format
WGBS	Whole Genome Bisulphite Sequencing
WGD	Whole Genome Duplication
WGS	Whole Genome Sequencing

Thesis structure

A description of the chapters making up this thesis is presented below, so that the reader may understand the reasoning behind including certain topics and writing them in this specific order.

Chapter 1 presents an introduction to ash trees and the threat of ash dieback disease; the reason for which the ash tree genome project was first initiated. I also describe recent research which found a low susceptibility phenotype present in a natural ash population. The reason for this chapter is to present the motivation behind sequencing the ash tree genome, which is the rapid response to a global health threat that can enable further comparative genomics research.

Chapter 2 is an introduction to current genome projects of forest trees, of which there are several. I also discuss the issues surrounding assembling heterozygous and/or very large genomes, which shaped the way in which we approached sequencing and assembling the ash tree genome.

Chapter 3 describes the methods and results of assembling the genome of a British ash tree. As this task was, necessarily, one of the first accomplished, it therefore makes sense to present these results first.

Chapter 4 follows on from the previous chapter to describe the separate assembly and annotation of the mitochondrial and plastid genomes. As these tasks used very different methods to the nuclear assembly, I think the reader would benefit from having these in a separate chapter.

Chapter 5 analyzes the history of whole genome duplications in ash, using the reference genome and transcriptome described in Chapter 3. As this method used only the reference sequences, I present these results after the two assembly chapters but before any other results which introduce additional data.

Chapter 6 is the first to use additional data; that of whole genome resequencing from 38 ash trees of diverse European origins. In this chapter I investigate population structure across these samples and estimate trends in historic effective population size.

Chapter 7 again uses new data, this time whole genome Bisulphite sequencing from seventeen *F. excelsior* and three *F. mandshurica* individuals. This results chapter is presented last, as the topic of epigenomics is notably different to that of genomics, on which all other chapters are focused. The data for this chapter were also analysed last chronologically.

Chapter 8 is a conclusion chapter that brings together all results and puts them in context of susceptibility to ADB. I describe other research that has already made use of my results, and discuss potential further work that could build on them, especially in areas that could aid the development of low susceptibility trees.

All tables and figures are placed as close to the referencing text as is reasonable, however some large tables are placed at the end of the chapter so as not to interrupt the rest of the chapter.

Contributions

Contributors to this thesis include:

- Chapters 3 and 4: Richard Buggs collected samples, Jasmin Zohren performed DNA and RNA extractions.
- Chapter 5: Laura Kelly obtained CDS for *Utricularia gibba* from the author of Ibarra-Laclette et al. (2013) and Endymion Cooper generated ORFs from transcriptome data for olive. Endymion also wrote the script RemoveRedundant.py which corrects for redundant Ks values. Gene annotation for ash was generated by Gemy Kaithokottil and David Swarbreck at The Genome Analysis Centre (TGAC, now the Earlham Institute).
- Chapter 6: DNA was extracted from 37 trees in the Realizing Ash's Potential (RAP) trial by Laura Kelly. DNA libraries were prepared and sequenced at TGAC.
- Chapter 7: Trees were grown at the University of Copenhagen by Erik Kjær, Lea Vig McKinney and Lene Rostgaard Nielsen. Lene Hasmark Andersen helped with DNA extraction. Bisulphite reaction and sequencing was carried out by Eva Wozniak at the Genome Centre, QMUL. Richard Nichols also contributed advice on statistics.

Associated publications

Portions of the work detailed in this thesis have been presented in national and international publications, as follows:

- Some parts of chapter 2 have been taken or modified from a book chapter 'Emerging Genomics of Angiosperm Trees' written by myself and supervisor Dr. Richard Buggs, as part of the book 'Evolutionary Genomics of Angiosperm Trees' editors Quentin Cronk and Andrew Groover, published by Springer. It is currently in press, but was published as an online pre-print on 31st December 2016, at: http://link.springer.com/chapter/10.1007/7397_2016_16. This chapter in included as Appendix 2 (page 170).
- Many of the results in chapters 3, 4, 5 and 6 have been published, albeit at much less detail, in the ash genome paper (Sollars et al. 2017, *Nature* Vol. 541, 212-216) on 12th January 2017. This article is included as Appendix 1 (page 147).
- Gene annotation results in Chapter 3 were used to aid the identification of expression markers associated with ADB. These results have been published in Harper et al. (2016), in which I am a co-author. Genes identified in this study as having expression associated with ADB susceptibility, were also used in Chapter 7.

Chapter 1

Introduction to ash trees and the threat from ash dieback disease

1.1 Biology and geography of *Fraxinus excelsior*

Fraxinus excelsior is a deciduous tree of the family Oleaceae, and one of ∼50 species within the *Fraxinus* genus (Fig. 1.1). It covers an extensive range across Europe; at its edges reaching to the UK, northern Spain, the Caucasus region, and southern tips of Scandinavia (Fig. 1.2)[EUFORGEN 2009]. It is found in a range of forest types, including coastal, hardwood, mixed, alluvial, and wetland forests [Marigo et al. 2000], as well as being a common hedgerow species [Thomas 2016]. European ash overlaps with the ranges of narrow-leaved ash (*F. angustifolia*) and manna ash (*F. ornus*) in southern Europe. It is thought that European ash, along with other deciduous trees, colonised northern Europe during the Cretaceous and gradually spread out and southwards during periods of cooling [Marigo et al. 2000]. Within the UK, *F. excelsior* is found across all areas of the country with the exception of the very north of Scotland (Fig. 1.3).

Historical data suggest that the European ash population expanded in the Holocene from numerous refugia. Both pollen and DNA data identify refugia in the Balkans, Carpathian mountains and Iberian peninsula [Thomas 2016; Magyari et al. 2014; Sutherland et al. 2010; Huertz et al. 2004a,b]. This is discussed in more detail in Chapter 6. Ash has been under decline for the past 5000 years, as found using pollen data, likely due to harvesting for wood and fodder during the Bronze Age. The 20th century saw a population expansion once again [Thomas 2016], possibly aided by the reduction in human rural populations [Marigo et al. 2000; Dobrowolska et al. 2011].

The gender structure of ash flowers is complex. Flowers can be fully female, fully male or hermaphroditic, though hermaphroditic flowers can sometimes have only one functional gender [Binggeli & Power 1991]. Ash trees can consist of all male, all female or all hermaphrodite flowers, or can be trioecious (a mixture of all three) [Binggeli & Power 1991; Morand-Prieur et al. 2003; Albert et al. 2013]. Hermaphrodite flowers are protogynous (change from female to male); though self-pollination is possible [Morand-Prieur et al. 2003], in natural populations ash is preferentially outcrossing [Thomas 2016]. Ash is wind-pollinated; pollen can travel large distances of up to 3km, though approximately 85% travel only up to 100m

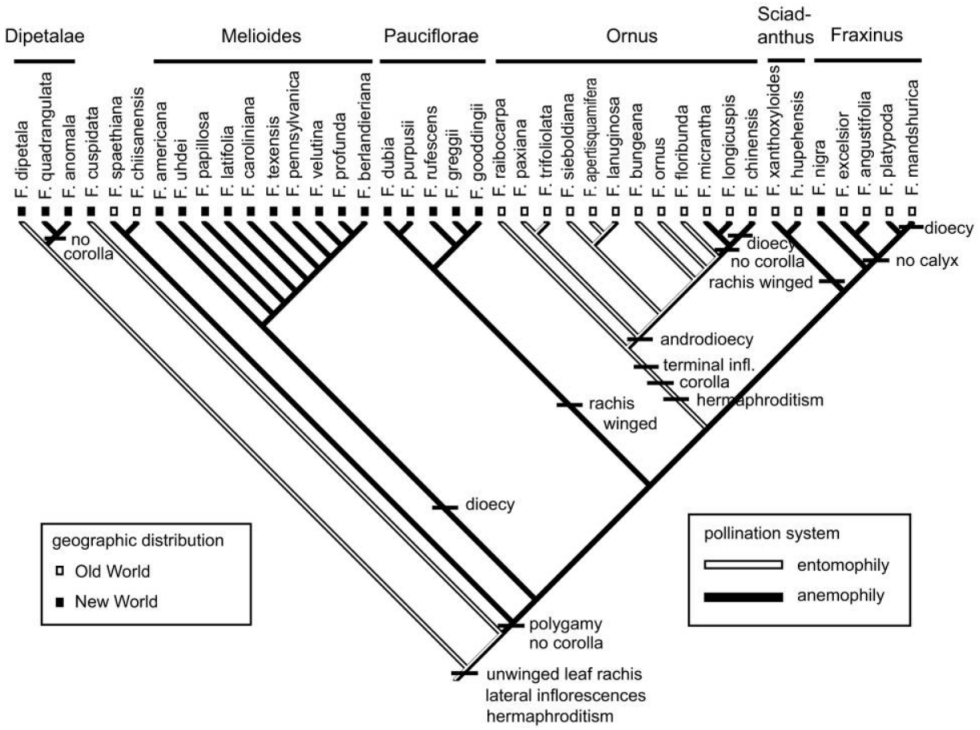


Figure 1.1: Phylogeny of *Fraxinus genus*, also showing geographic distribution, pollination system and reproductive system. *F. excelsior* lies within section Fraxinus. Image taken from Whitehill et al. (2011).

[Bacles et al. 2005; Bacles & Ennos 2008]. Hybridisation with *F. angustifolia* occurs readily where their ranges overlap [Gerard et al. 2013; Huertz et al. 2006). Seed production begins in the tree's 20-30th year [Dobrowolska et al. 2011]. Seeds develop through July to September [Thomas 2016]. They are held in a flattened wing (samara) and form clusters that hang in bunches. The seeds rotate when falling (between September and March) and depending on wind speed, can reach distances of 100m from the mother tree [Dobrowolska et al. 2011].

Ash is easily established, building an extensive root and shoot system quickly after establishment [Kerr & Cahalan 2004]. As a shade-tolerant species, seedlings employ a 'gap species' growth strategy; they can change growth rate depending on light availability, so that in deep shade, growth will be reduced [Petritan et al. 2007; Thomas 2016]. This serves to reduce unnecessary investment in growth with small gain in light availability, therefore allowing ash to survive long periods in shade. It also shows plasticity in crown topology by displaying most of the leaf area at the top of the crown to minimize self-shading and to enhance light interception [Petritan et al. 2007]. Ash grows optimally in warm climates, with moist soil that is rich in nutrients. It has an intermediate tolerance to flooding, though prolonged water-logging is detrimental [Kerr & Cahalan 2004; Glenz et al. 2006; Dobrowolska et al. 2011]. Growth is also impacted by low foliar nitrogen levels [Kerr & Cahalan 2004; Dobrowolska et al. 2011]. Being tolerant of drought, the range of ash is expected to expand given future climate change predictions, without any other limiting factors [Thomas 2016; Goberville et al. 2016].

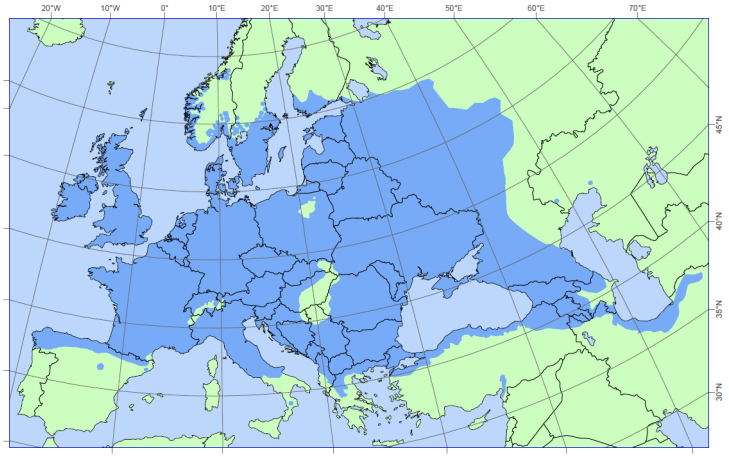


Figure 1.2: European distribution of *Fraxinus excelsior*, with current range shown in blue areas. Image downloaded from EUFORGEN webpage http://www.euforgen.org. Last accessed 23rd April 2017.

1.2 Value of ash

As a wood used primarily in furniture and construction, the commercial value of ash in the UK is estimated to be £22 million per annum [DEFRA 2013]. The ash supply chain is complex as there are many different uses for both the trees and its wood. Ash wood is used for construction, furniture, crafts, as fuel wood, in biomass generators, and to make charcoal. Trees are sold for landscaping purposes, e.g., along roads and motorways, in private gardens, as well as for recreational purposes such as in outdoor centres and botanical gardens. Seeds and seedlings are also traded by garden centres and nurseries to provide to the landscaping business.

Further monetary values can be placed upon non-commercial services, such as to the environment and society. Ash trees in the UK are estimated to sequester 0.7-1.0 million tonnes of carbon dioxide per year, which can be valued between £41-58m [DEFRA 2013]. Other environmental services include air and water quality improvement, shade creation, wind control, pollution and noise reduction, and flood alleviation [DEFRA 2013]. Ash also supports a wide range of biodiversity, being a key ecological forest species; it hosts almost one thousand associated species and fourty-four obligate species [Mitchell et al. 2014]. The DEFRA 2013 report places an estimated figure of £150m on the combined social and environmental uses of ash. In view of the large value of ash in the commercial wood sector and even larger value supporting social and environmental services, implications clearly arise if ash were to be removed from the country's woodlands.

1.3 *Hymenoscyphus fraxineus*, the causative agent of ash dieback

European ash has been under attack from the fungal pathogen *Hymenoscyphus fraxineus* over the two past decades. The pathogen is the cause of the ash dieback (ADB) disease [Kowalski 2006, Kowalski & Holdenreider 2009, Queloz et al. 2011; Zhao et al. 2012; Baral et al. 2014]; an infection characterised by black lesions on stems and branches, defoliation (the wilting or loss of leaves), leading to a reduction in photosynthesis and necrosis of the

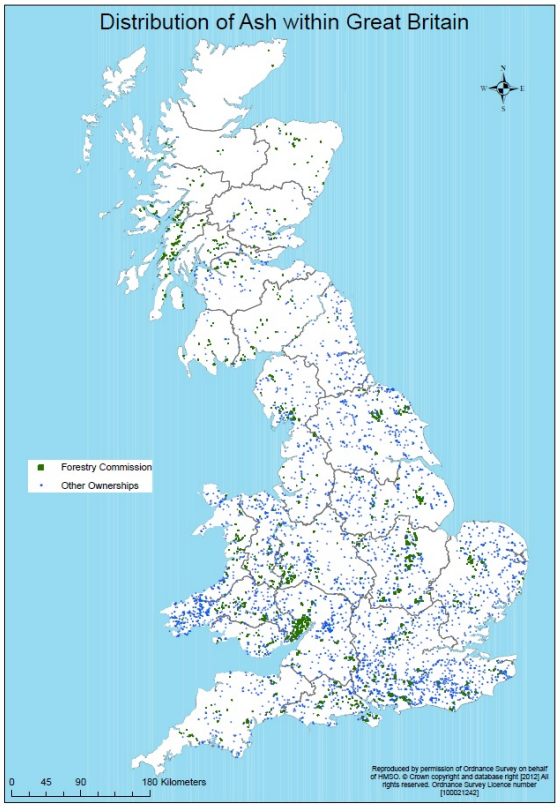


Figure 1.3: Distribution of ash within Great Britain. Green squares indicate Forestry Commission sites, and blue dots show other ownership. Image last downloaded from Forestry Commission www.forestry.gov.uk on 1st November 2016.

xylem vessels, thus reducing its ability to transport water [Tulik 2010]. The infection also renders the tree increasingly vulnerable to additional infections. Together, these ultimately lead to the death of the tree. The life cycle of *H. fraxineus* is shown in Figure 1.4: spores infect the tree via the leaves or roots, and fall back onto the ground as fruiting bodies in infected leaves where they can propogate further. From the soil, mycelium can infect the roots of mature trees or saplings. Asexual or sexual sporulation can also occur, releasing spores into the air where they can infect the leaves of other ash trees.

Ash dieback is thought to originate from Asia [Zhao et al. 2012, Gross et al. 2014] where a related species to *F. excelsior*, Manchurian ash (*F. mandshurica*), is naturally resistant to the disease due to co-evolution of host and pathogen. The disease was first identified in Europe in the 1990's in Poland [Kowalski 2006] from where it has since spread throughout the rest of Europe, reaching Scandinavia in the early 2000's [Timmermann 2011] and Western Europe later that decade [Ioos 2009] (Figure 1.5). It was also found to infect *F. excelsior*'s European sister species, *F. angustifolia* [Kirisits 2010]. Most recently in 2012, it was identified in the UK, likely transported over from mainland Europe along with the import of infected trees. As of 3rd October 2016, there are 1,037 10km grid squares across the UK with one or more confirmed sites of ADB infection (Figure 1.6, forestry.gov.uk/chalara), equal to 36.6% of all 10km grid squares in the UK.

In Denmark, an estimated 90% of ash trees have been affected by ADB [McKinney et al. 2012], and likewise an estimated 90-99% of UK trees are at risk of infection [Cormier 2012; Denman & Webber 2013]. The UK government has estimated losses of up to £50 million

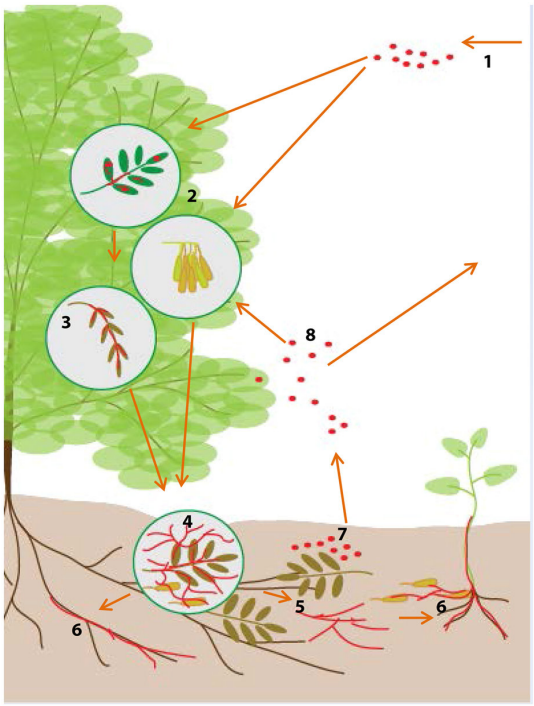


Figure 1.4: Life cycle of the fungal pathogen *Hymenoscyphus fraxinues*. Spores carried on the wind (1) infect ash trees through the leaves or seed cases (2). Fruiting bodies form on the surface of leaves or rachis (3). Infected leaves fall to the ground, allowing the fungus to propogate further in the leaf litter (4) and in the soil itself (5). Fungal mycelium can then infect the roots of mature trees or saplings (6), and sporulation (7) can release asexual or sexual spores into the air (8) where they can re-infect the host tree or can be carried on the wind to infect others. Image taken from Fones et al. (2016)

per annum [DEFRA 2013] due to ADB. This is split into two categories: an estimated £22m from commercial sectors (e.g., loss of wood for use as timber e.g., in construction, furniture) and £38m from social / environmental sectors, such as for recreation, landscape, biodiversity support, carbon sequestration and air pollution absorption. Ash also hosts a small ecosystem of fourty-four obligate species (made up of 11 fungi, 29 invertebrates, and 4 lichens), 62 highly-associated species and nearly a thousand further associated species [Mitchell 2014]. *F. excelsior* cannot be replaced by another single tree species. However many of its functions and associated species can be offset by other common forest species such as oaks (*Quercus robur* and *Q. petraea*) and aspen (*Populus tremula*). Considering both the economic and ecological value of ash, a strong case can be made for funding research into the effects of the disease, its projected spread and potential solutions.

In addition to ADB, there is a potential threat from the US of the wood-boring beetle Emerald Ash Borer (EAB, *Agrilus planipennis*). Similar to ADB, the EAB was accidentally introduced from Asia [Herms & McCullough 2014]. First discovered in the US in 2002 [Cappaert et al. 2005], it currently infects various *Fraxinus* species in the North Eastern states and is gradually spreading outwards [USDA 2016]. Adult beetles lay eggs in bark crevices, which upon hatching, burrow internally through the tree to reach the xylem, phloem and cambium tissues. After six developmental stages (four larval instars, prepupae and pupae), the new adults then bore back through the bark to exit the tree [Herms & McCullough 2014]. There is concern that the beetle could also be introduced to Europe where any ash populations surviving the ADB infection or escaping exposure, could be vulnerable to a second threat of the EAB [www.forestry.gov.uk/emeraldashborer].

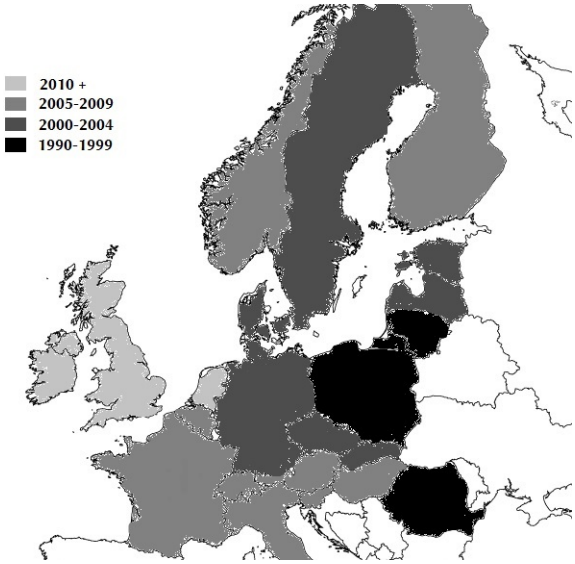


Figure 1.5: Year of first identification of ash dieback in European countries, from oldest incidence (black) to most recent (light grey). (Original image).

1.4 Natural genetic variation in susceptibility to ADB

Fortunately, a small percentage (around 1%) of trees are found to have natural low susceptibility to ADB [McKinney et al. 2011; Kjaer et al. 2012; McKinney et al. 2012]. Trials in Denmark involved disease symptom assessment in 39 clones selected from native *F. excelsior* trees around the country. On average, 50 replicates of each genotype were generated using grafting, and were established at two sites in Denmark. Crown damage was assessed over a three year period, as well as stem necroses over two years. The genetic variation involved in the degree of damage observed was highly significant, and the genotypic correlation was very high between the two test sites. Particular genotypes showed fairly consistent low susceptibility to ADB. Broad-sense heritability was calculated at around 0.4 for the two sites across three years [McKinney et al. 2011]. One particular genotype, 'clone 35', was found to have very low susceptibility to ADB, consistent across the majority of its clones. Additional tests in a controlled greenhouse environment confirmed the role genetics plays in the phenotype, and also suggested a mechanism for the observed low susceptibility. McKinney et al. (2012) found that low susceptibility trees can limit the growth of fungal necroses when inoculated, suggesting that the low susceptibility arises through a form of inhibition rather than a tolerance.

The implications of finding low susceptibility genotypes depend on the level of additive genetic variation contributing to the phenotype, i.e. how well the phenotype can be maintained or improved upon during selection. Kjær et al. (2012) found high levels of additive variation, calculating narrow-sense heritability between 0.37 and 0.52. This means that the low susceptibility trait will likely respond well to selection and improvements in resistance could be gained by breeding low susceptibility trees.

Genetic variation in susceptibility has also been found in other tree species, for example in European elm species susceptible to pathogenic species of *Ophiostoma* (causative agents of Dutch elm disease) [Martin et al. 2013], and in American chestnut susceptible to blight

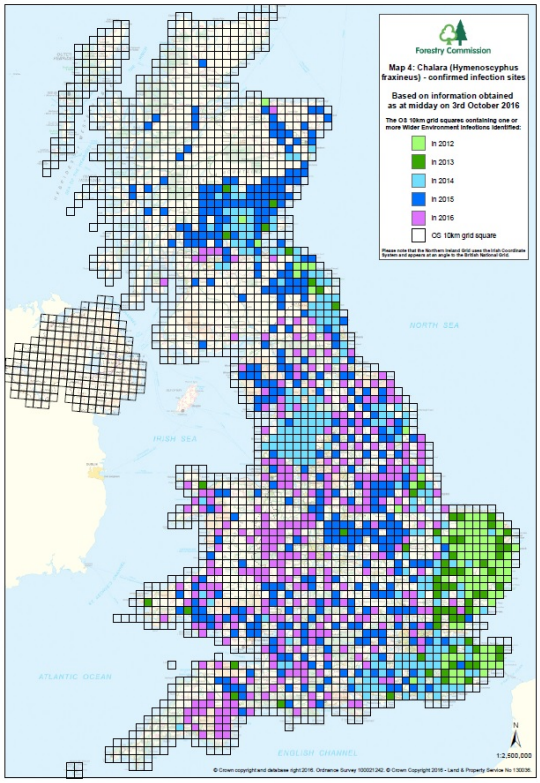


Figure 1.6: Distribution of confirmed ADB sites across UK, as of 3rd October 2016. Image downloaded from www.forestry.gov.uk/chalara on 30th October 2016

fungus [Griffin 2000]. The role that genetics seemingly plays in ADB susceptibility means that changes in the DNA sequence (e.g., nucleotide polymorphisms, structural variants), or changes in the levels of gene expression, or a combination of the two, could be associated with the phenotype.

Clone 35, along with several other genotypes, have now been developed as resources to study the genetic differences behind ADB resistance [Harper et al. 2016], some of which are used in this thesis in Chapter 7. Using a panel of 213 Danish *F. excelsior* trees, Harper et al. (2016) were able to find Gene Expression Markers (GEMs) that were significantly associated with susceptibility. The genes identified were largely made up of MADS-box proteins and transcription factors. Sollars et al. (2017) screened a refined list of these markers in a panel of British trees, finding somewhat lower susceptibility in general among the British trees tested in comparison to the Danish trees (performed by Andrea Harper & Ian Bancroft, University of York). Further metabolomic work identified a group of iridoid glycosides to be highly associated with the low susceptibility phenotype (performed by Christine Sambles & Murray Grant, University of Warwick).

Using the identified low susceptibility trees as a breeding resource could help to rebuild ash populations that can inhibit the growth of the pathogen. Genetic testing for the variants or expression markers associated with the trait can greatly speed up breeding low susceptibility trees by using marker assisted selection methods. The genome sequence has already proven a valuable resource in identifying some of these expression markers, and will continue to be so in future studies.

1.5 Current status of ADB research and future directions

A large-scale trial is being carried out in the UK and the US to find genes associated with resistance to both ADB and EAB within the whole ash genus (Tree Health and Plant Biosecurity Initiative funded project, led by Richard Buggs at QMUL). Some *Fraxinus* species are already known to display varying levels of susceptibility to ADB, for example *F. ornus*, *F. angustifolia*, and *F. pennsylvanica* all show susceptibility but less so than *F. excelsior* [Kirisits & Schwanda 2015; Schwanda & Kirisits 2016; Gross & Sieber 2016], while *F. mandshurica* and *F. americana* are known to have very low susceptibility [Denman & Webber 2013; Gross et al. 2014; Zhao et al. 2012]. The North American species *F. americana* and *F. pennsylvanica* show susceptibility to EAB, while *F. mandshurica* is again resistant [Rebek et al. 2008]. However, the susceptibility of many other ash species to either ADB or EAB is not yet known, particularly the species residing in regions where the threat has not yet reached (i.e., European species to EAB, and American species to ADB). A paper recently published by the University of Copenhagen group, tested seventeen *Fraxinus* species for levels of susceptibility against ADB [Nielsen et al. 2016]. They found that susceptibility followed a phylogenetic pattern as opposed to a geographical paterrn; for example, five species tested from section Ornus displayed few symptoms, whereas the four species of section Fraxinus displayed more severe symptoms (Fig. 1.7).

To test the level of susceptibility to both threats, multiple genotypes from 31 ash species will be inoculated with EAB in the US, with a replicated trial of ADB susceptibility in the UK. For the ADB trial, more than 3000 grafts have been attempted for over 120 genotypes [L. Kelly, pers. comm.]. The project aims to identify the genes involved in low susceptibility by building phylogenetic trees for thousands of genes across the sequenced ash species. By utilising the susceptibility data of the 31 species, genes that generate a phylogenetic tree with a similar topology to the susceptibility data can be identified as potential resistance candidates [http://www.bbsrc.ac.uk/funding/filter/tree-health-and-plant-biosecurity-phase2/]. This method has previously been used to identify genes involved in parallel evolution of echolocation in mammals [Parker et al. 2013]. Whole genome sequencing has already been performed on trees covering many of the 31 species. Assembly of these is currently being aided by the reference genome sequence of *F. excelsior* [L. Kelly, pers. comm.]. A potential challenge in using this topology comparison method could be the quality of gene assembly within the genome sequence of each species. Accurate alignment of gene sequences, and therefore phylogenetic tree topology, could depend on the completeness on the gene sequences. Any non-contiguous sequences may be missed from an alignment which could skew the resulting gene tree topology. Using the *F. excelsior* reference sequence to guide assembly could help to assemble gene content in other *Fraxinus* species, though it should be noted that the *F. excelsior* assembly is also not complete. Another challenge could be in the method of phenotyping susceptibility to ADB and EAB in the trees, and the particular measure by which susceptibility is assessed (e.g., levels of visible damage, levels of pathogen detected within tree). It may also be difficult to associate this continuous measure with a tree topology that splits the genus into discrete groups.

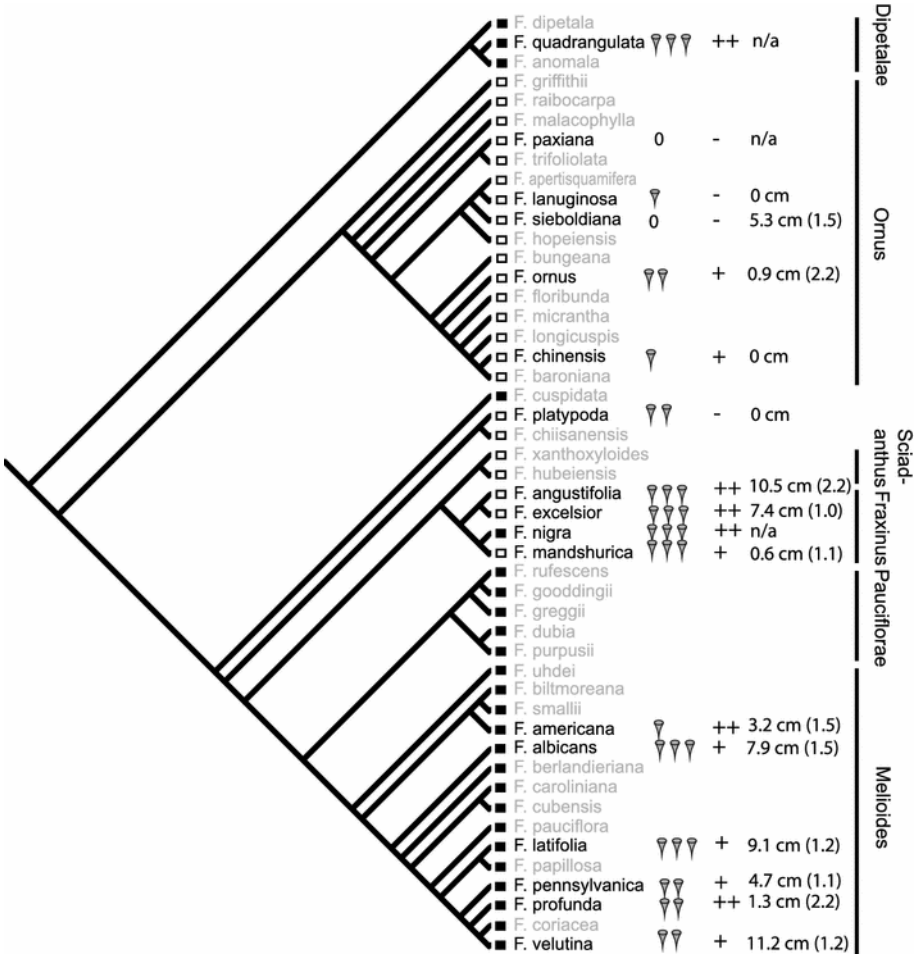


Figure 1.7: Phylogenetic overview of *Fraxinus* species tested in Nielsen et al. (2016) (black) summarizing results of 1) apothecium formation on rachices of overwintered leaves, 0: no apothecia, 1: At least one sample with 1-3 apothecia, 2: At least one sample with 4-10 apothecia, 3: At least one sample with >10 apothecia; 2) Crown damage observed on trees in the arboretum, no symptoms, + few symptoms, <10% damage, ++ >10% damage; 3) Average lesion length (standard error) after controlled stem inoculation. Image taken from Nielsen et al. (2016)

Another project in the same funding stream, led by James Brown at the John Innes Centre, will investigate the ecological genetics of the ADB pathogen, *Hymenoscyphus fraxineus*. The project will investigate the genetic diversity and pathogenicity of the fungus in different trees and across populations, as well as comparing the pathogen with its non-virulent sister species, *H. albidus*. Using this information, the project aims to identify ways to manipulate the pathogenicity of the fungus and therefore increase survival among ash trees.

Trials conducted by Forest Research are aiming to identify 500 trees with low susceptibility [Jo Clark, pers.comm.]. These could provide a pool from which to breed low susceptibility progeny. The high number of trees identified will ensure that genetic diversity is preserved throughout the genome and will minimise the effect of inbreeding in later generations. It is imperative to maintain a level of genomic and phenotypic variation so that the population

can adapt to further biological or environmental stresses.

In the face of the threat from ADB, identifying and breeding low susceptibility genotypes will be key to the regeneration of ash populations across Europe. There have already been successful studies that find both low susceptibility trees and markers associated with the trait [Harper et al. 2016; McKinney et al. 2011]. However the sexual reproduction of the pathogen may allow it to overcome the small window of low susceptibility in the host [Landolt et al. 2016]. Recommendations involve collecting germplasm from all trees identified as having low susceptibility to preserve the genetic information in these genotypes and establish a breeding program [Pautasso et al. 2013]. These valuable trees should also be protected from competition with other species and from other environmental stresses, in order to maximise their survival [Pautasso et al. 2013]. Re-planting of ash will need to take into account the characteristics of site requirements for optimal seedling growth, such as soil moisture and nutrient content. In a population already stressed by infection, steps should be taken to minimise environmental stress by planting trees in optimal sites.

In conclusion, a genome sequence would be a valuable resource to any future studies on ash genetics. This includes potential breeding trials for low susceptibility trees to identify genomic markers associated with ADB susceptibility, as well as comparative genomics and phylogenomics studies using other *Fraxinus* species. The use of genomic methods allows rapid selection of low susceptibility progeny, and can greatly accelerate the generation of resistant trees. Re-stocking natural woodlands with these trees could be an effective defence against the threat of ADB.

Chapter 2

Introduction to current genome projects of forest trees

Genome sequencing of trees has lagged somewhat behind that of herbaceous species. The first tree genome, *Populus trichocarpa* [Tuskan et al. 2006], was published six years after the first plant genome [Arabidopsis Genome Initiative 2000]. According to my estimation, there are currently 61 non-woody plant genome sequences either published or available in the Phytozome v11.0 resource, compared to only 31 tree or woody species, including those listed in Table 2.1, as well as the olive tree genome. This number of 31 can be further split up into 19 fruit, seed or nut trees, compared to 12 forest tree species. Although my estimations have likely missed many herbaceous genomes released in other repositories, but possibly fewer tree genomes, it is clear that there is an enrichment in herbaceous plant genome sequences available in comparison to trees, and an enrichment of crop trees in comparison to forest trees. One reason for this is that there are breeding programs established for many crop plants (cultivated trees included), in which breeding new varieties gains massively from using genetic information. Therefore there is both the motive to sequence the genome, in order to provide a reference for cultivar comparisons and enable marker-assisted or genomic selection, as well as the means from the commercial interest invested in the genome project. In practical terms, trees also take longer to grow and reach maturity than herbaceous species, therefore experimenting to generate new genotypes takes a great deal longer and much more space than a small plant that produces seeds in just a few weeks.

Many tree genome projects are now underway (see Section 2.2), including forest trees that have either ecological importance or have industrial interest as timber crops. There still remain several challenges; to achieve good quality DNA from plants, sequence it successfully, and assemble it into a contiguous, useful genome sequence. Though these have not prevented genome projects from being carried out, they are certainly considerations to be taken into account when developing a sequence and assembly strategy. These challenges will be explained in this section. Section 2.2 will then describe some current forest tree genome projects that I feel are relevant; either they have a particular interesting reason behind being sequenced in the first place (such as biomass crops, disease resistance, insights into forest ecology), or they have overcome the challenges mentioned using optimised sequencing or assembly methods. I also present summarised information of tree genome projects in Tables 2.1 and 2.2. Much of this section has been modified from a book chapter written by myself

and supervisor Richard Buggs (see Associated Publications and Appendix 2), and this is noted at the start of each relevant subsection. In addition, I detail some online databases from where large collections of tree genomic data and information can be accessed.

2.1 Challenges in sequencing and assembling plant genomes

2.1.1 Heterozygosity

Many plant species are outcrossing; compatibility upon fertilisation requires individuals to have different genotypes. Various self-incompatibility mechanisms exist to prevent pollen fertilising the stigma of the same individual, for example, some trees produce only all male or all female flowers, whereas others involve biochemical mechanisms. Outcrossing exists to increase genetic diversity; a diverse population will possess sufficient variation to evolve quicker in changing environments. Many plant species are also wind-pollinated, meaning that pollen often travels long distances, leading to one or a few extensive, genetically diverse populations instead of lots of small localised populations. The constant mixing of genetic material leads to increased heterozygosity in the genome; sites that possess different alleles. The human population experienced a bottleneck approximately 50,000 years ago [Schiffels & Durbin 2014; Henn et al. 2012] leading to loss of heterozygosity and a genome with relatively infrequent polymorphisms. In contrast, plant (and many animal) genomes have a much higher frequency of polymorphisms that lead to difficulties in assembling the genome computationally.

Short reads from Next Generation Sequencing (NGS) platforms are assembled into a reference using de bruijn graphs. These graphs use the overlaps in sequences to generate one longer consensus sequence. This is explained in more detail in Section 3.1.1. However when polymorphisms occur in the reads, there becomes more than one possible path through the graph, one with each allele. With increasing densities of polymorphisms, the graph becomes very complex, and deciding the correct consensus sequence is sometimes not possible. Therefore the sequence can be split, retaining each haplotype in different contigs as if they were paralogs rather than alleles. Although this keeps all haplotype information in the assembly, mapping reads back to the reference becomes a problem when near identical contigs are present, as they will map equally well to more than one place. Ideally, alternative haplotypes should be condensed into one reference, and all reads (which still contain the haplotype information) can be mapped to the one 'master' copy and then phased into haplotypes later. However very diverse regions of the genome are sometimes impossible to resolve into a single reference.

Many commercial crop plants have been repeatedly crossed to generate inbred lines, which aids genome sequencing by producing long stretches of homozygosity. These resources would take a lot of time and money to generate in most trees, due to their long generation time. In the case of ash however, a progeny of a selfing experiment performed on an already low-heterozygosity tree was available (explained further in section 3 which describes the assembly of the reference tree).

2.1.2 Large genome size

Smaller genomes tend to be favoured for any model organisms as they are easier and cheaper to sequence, especially when multiple individuals or genotypes are involved. Although there are many herbaceous plants with fairly extensive genomes (e.g. the largest plant genome known so far is found in the herbaceous species *Paris japonica* [Pellicer et al. 2010]) the challenges of sequencing large genomes are particularly relevant in trees, as the genomes of some gymnosperms have reached upwards of 20 Gbp [Birol et al. 2013; Neale et al. 2014]. These very large genomes occur through processes such as Whole Genome Duplications (WGD) [Soltis & Soltis 2013] or hybridisation causing polyploidy (see chapter 5), proliferation of transposable elements [Michael 2014; Morse et al. 2009], and unequal recombination of repetitive elements. Alternatively, Federoff [2012] hypothesises that it is the epigenetic silencing of transposable elements, preventing their mobility and leading to unequal recombination, that has caused many genomes to reach large sizes.

Sequencing and assembling large genomes poses many problems. Firstly, as the amount of DNA needed to be covered increases, so does the number of libraries, flow cells, and sequencing runs required, as well as the amount of data produced. This consequently increases the costs of laboratory equipment, consumables, sequencing, and data storage. In addition, a large amount of computing power is required to process all reads produced and assemble the genome sequence, therefore extra costs can be incurred from having to buy new powerful computers with large amounts of RAM (Random-Access Memory). The time needed to process all reads from initial transfer and QC, to assembly, mapping and variant calling can be considerable, especially if parameters need to be continually optimised or methods repeated. Available software may also inefficiently handle a large number of reads. For example, the genome project of white spruce found that new assembly algorithms were required to deal with the large amount of data [Birol et al. 2013].

The repetitive nature of large genomes also presents difficulties upon assembly; this is explained more thoroughly in section 3.1.1. In brief, genomes are assembled from the overlap of short sequenced reads, or even shorter k-mers. In repeat regions, many k-mers contain the same sequence while originating from different regions of the genome. The assembly process cannot distinguish them and often collapses these regions into one, leading to reduced assembly sizes and broken contigs. If these regions are separated out, read mapping can then become problematic as there will be large numbers of reads with non-unique mapping locations. One solution is to use longer read lengths that have a greater chance of spanning the repeat region; then repeats can be assembled and the reads will likely map uniquely to the correct region.

2.2 Current forest tree genome projects

Many tree genomes have now been sequenced or have genome projects underway, in part due to the continuous reduction of sequencing costs and development of bioinformatic tools. The first tree genome sequence, *Populus trichocarpa*, was published ten years ago [Tuskan et al. 2006]. Since then, tree genome projects have been focused largely on commercially

important species; usually fruit and nut trees such as apple [Velasco et al. 2010], peach [Verde et al. 2013], walnut [Martinez-Garcia et al. 2016], European hazelnut [Rowley et al. 2012], cacao [Argout et al. 2011], coffee [Denoeud et al. 2012] and several citrus trees [Wu et al. 2014; Xu et al. 2013]. In addition, the genomes of palm species *Phoenix dactylifera* (date palm) and *Elaeis guineensis* (oil palm) have been assembled and published [Al-Dous et al. 2011; Al-Mssallem et al. 2013; Mathew et al. 2014; Singh et al. 2013]. Very recently, the olive tree genome has been published [Cruz et al. 2016].

Although many forest tree species are commercially viable as timber crops, there are significantly fewer forest tree genome sequences available. The following sections of this chapter will give an overview of the genomic data available for both angiosperm and gymnosperm forest trees, and of large multi-species sequencing projects currently underway. Table 2.1 lists forest tree genome sequences currently available, and Table 2.2 lists genome-wide datasets available for many other angiosperm forest tree species.

Table 2.1: Forest tree species with genome sequence available. Blank cells represent that the information cannot be found, or assemblies do not have a version number (presumably at first version). References: [1] Tuskan et al. (2006), [2] popgenie.org. [3] Nystedt et al. (2013), [4] Birol et al. (2013), [5] Warren et al. (2015), [6] Neale et al. (2014), [7] Zimin et al. (2014), [8] hardwoodgenomics.org, [9] oakgenome.fr, [10] phytozome.jgi.doe.gov, [11] Dai et al. (2014), [12] Wang et al. (2013).

Species	1C genome size (Mbp)	Assembly version	No. scaffolds	Assembly size (Mbp)	Scaffold N50 (kbp)	No. genes annotated	Ref
Populus trichocarpa	485	v3.0	1,446	422.9	19,500	41,335	[1]
Populus tremuloides		v1.1	164,504	377	15.2	35,694	[2]
Populus tremula		v1.1	204,318	386	43.9	34,152	[2]
Picea abies	19.6 Gbp	v1.0		12 Gbp	4.9	28,354	[3]
Picea glauca	20 Gbp	v4.0	4.3 million	22.5 Gbp	20.9	16,386(v3)	[4],[5]
Pinus taeda	22 Gbp	v2.0		22,104	107.8	50,172(v1)	[6],[7]
Castanea mollissima	794	v1.1	41,260	724	39.6	36,478	[8]
Quercus robur	740	v1.0	18,000 >2000 bp	1350	257	54,000	[8],[9]
Salix purpurea	450	v1.0	7,528	392	17,359	37,865	[10]
Salix suchowensis	429		103,144 >=100 bp 7,516 >= 2,000 bp	304	925	26,599	[11]
Betula nana	450		551,923	564	18.79	none	[12]

2.2.1 The first tree genome; *Populus trichocarpa*

Four years after the publication of the *Arabidopsis thaliana* genome sequence, and many other crop plants, the first tree genome was released [Tuskan et al. 2004], and published two years later [Tuskan et al. 2006]. *Populus trichocarpa* was chosen as a model species for woody plants due to its relatively small genome size (485 Mbp), rapid growth, and availability of genome-wide data such as ESTs, genetic maps and BAC libraries at the time [Bradshaw et al. 2000; Tuskan et al. 2004]. In fact, many Quantitative Trait Loci (QTL) for traits such as growth and wood form had already been mapped (e.g. Bradshaw & Settler 1995), albeit at low resolution. It also represented the first perennial plant to be sequenced [Tuskan et al. 2004], with the intention that by comparing the genome to that of annual *Arabidopsis*

Table 2.2: Angiosperm forest tree species with genome-wide data available. WGS: Whole Genome Sequencing (data), SSRs: Small Sequence Repeats, EST: Expressed Sequence Tags, GBS: Genotyping by Sequencing, RAD-seq: Restriction site-Associated DNA sequencing. URLs: [1]hardwoodgenomics.org [2]fagaceae.org [3]birchgenome.org [4]oakgenome.fr [5]w3.pierroton.inra.fr/QuercusPortal/ [6]phytozome.jgi.doe.gov/pz/portal.html [7]popgenie.org [8]willow.cals.cornell.edu/genome/ [9]https://www.ncbi.nlm.nih.gov/bioproject/203514

Tree	Species	Lead Researcher(s) / Group	Data available	URL
Ash	*Fraxinus pennsylvanica*	Hardwood Genomics	WGS, SSRs, RNA-seq	[1]
	Fraxinus americana		WGS, SSRs	
Beech	*Fagus grandifolia*	The Fagaceae Project	RNA-seq, EST assembly	[2]
Birch	*Betula nana*	Richard Buggs, QMUL	WGS, genome assembly, RAD-seq	[3]
	Betula platyphylla	Chunping Yang, Northeastern Forestry University	WGS, genome assembly, gene annotation	
Black Cherry	*Prunus serotina*	Hardwood Genomics Project	WGS, SSRs	[1]
Black Walnut	*Juglans nigra*	Hardwood Genomics Project	WGS, SSRs, RNA-seq, ddRADtag (ongoing)	[1]
Blackgum	*Nyssa sylvatica*	Hardwood Genomics Project	WGS, SSRS, RNA-seq	[1]
Chestnut	*Castanea crenata*	The Fagaceae Project	EST assembly	[2]
	Castanea dentata		EST assembly	
	Castanea mollissima		WGS, EST assembly, Physical map	
	Castanea sativa		EST assembly	
Honeylocust	*Gleditsia triacanthos*	Hardwood Genomics Project	WGS, SSRs, RNA-seq,GBS (ongoing)	[1]
Oak	*Quercus alba*	Hardwood Genomics Project	WGS, SSRs, EST assembly	[1]
	Quercus robur	Christophe Plomion, INRA	WGS, genome assembly, SNPs, transcriptome assembly, genetic map	[4,5]
	Quercus rubra	Hardwood Genomics Project	WGS, RNA-seq, ddRADtag (ongoing)	[1]
Poplar / aspen	*Populus trichocarpa*	Poplar Consortium	WGS, genome assembly & annotation, transcriptome, physical & genetic maps, RNA-seq	[6]
	Populus tremuloides	UPSC	WGS, genome assembly & annotation	[7]
	Populus tremula	UPSC	WGS, genome assembly & annotation	[7]
Redbay	*Persea borbonia*	Hardwood Genomics Project	WGS, SSRs	[1]
Sugar Maple	*Acer saccharum*	Hardwood Genomics Project	WGS, SSRs, RNA-seq	[1]
Sweetgum	*Liquidambar styraciflua*	Hardwood Genomics Project	WGS, SSRs, RNA-seq	[1]
Tulip poplar	*Liriodendron tulipifera*	Hardwood Genomics Project	RNA-seq, GBS (ongoing)	[1]
Willow	*Salix purpurea*	Larry Smart, Cornell University	WGS, genome assembly, gene annotation	[8]
	Salix suchowensis	Tongming Yin, Nanjing Forestry University	WGS, genome assembly, EST assembly	[9]

thaliana, genes specific to, or enriched in, long-lived plant species could be identified. This could indicate some evolutionary adaptations of woody plants to their life history, such as secondary xylem (wood) formation, complex crown formation, long-lived host-parasite interactions as well as an overall larger size [Bradshaw et al. 2000; Tuskan et al. 2004].

The *Populus* genome project was initiated in May 2002 by the US Department of Energy [Tuskan et al. 2004]. Sequencing was carried out using a shotgun approach with paired reads from a 3kb and 8kb library, as well as using BAC and fosmid libraries to order contigs. Further, the availability of genetic and physical maps linked by Simple Sequence Repeat (SSR) markers, helped to assign and order scaffolds on to chromosomes. The original assembly consisted of 2,447 scaffolds in 410 Mbp, with 355 Mbp assigned to chromosome-scale linkage groups [Tuskan et al. 2006]. The collection of EST data also aided gene annotation, of which 45,555 were annotated. 12% of genes (5,428) shared no ortholog in the *Arabidopsis* genome, representing uniquely sequenced genes. Genes in *Populus* were enriched for functions such as insect and disease resistance, meristem development, lignin and cellulose biosynthesis and nutrient transport in comparison to *Arabidopsis*, confirming the hypothesis that these types

of genes would be enriched in long-lived woody species.

The genome sequence was further improved on two occasions. The most recent version benefits from a high density genetic map which allowed the main genome sequence to be arranged in 1,446 scaffolds. Researchers also attempted to merge outbred haplotypes, encompassing more genetic information from the *Populus* population as a whole. Additional RNA-seq data coupled with EST assemblies from related *Populus* species enabled an improved transcriptome annotation, resulting in 41,335 annotated genes.

The genome assemblies of *Populus trichocarpa* demonstrate a long-term continuous effort to improve the reference sequence of the model tree species. The genome has laid foundations for further studies on comparative and evolutionary genomics of trees, as well as given an experimental background for genetic improvement and commercial breeding.

2.2.2 Large and complex genomes - Gymnosperms *Pinus taeda* and *Picea* spp.

Reaching upwards of 20 Gbp, the genomes of gymnosperms are among the largest ever sequenced. Assembling genomes of this size produces significant challenges, as mentioned previously. Here I will compare the sequencing and assembly strategies of three gymnosperm species, loblolly pine (*Pinus taeda*), white spruce (*Picea glauca*) and Norway spruce (*Picea abies*), which have all been optimised in various ways to handle the large genome size. Although the assemblies are not complete, they have all been annotated with a number of genes comparable to other trees (Table 2.1), suggesting that the gene content at least seems to have assembled well. The remaining genome elements, being made up largely of repeats, are notoriously difficult to assemble and therefore may be collapsed or missing in these current assemblies.

The team behind the assembly of white spruce [Birol et al. 2013] decided to employ a shotgun assembly strategy, noting the saving in cost compared to localised sequencing protocols such as fosmid libraries, or long read technology such as Moleculo. The saving is especially substantial with such a large genome size. The genome was sequenced on Illumina HiSeq 2000 and MiSeq (the MiSeq platform was modified to generate longer read lengths of 500 bp), which therefore allowed a combination of high coverage and long read data. A range of insert size libraries were used; between 250 bp and 12 kbp, with a total genome coverage of 67x. The reads were assembled using ABySS [Simpson et al. 2009] with some modifications to account for the huge size of the dataset. The advantage of running ABySS is that the program is modular, and can therefore be run in separate parts with optimisation at each step, without needing to load all data and produce results in one go. The assembly was later improved [Warren et al. 2015], by re-scaffolding using an RNA-seq transcriptome assembly, cDNA clone sequences and additional mate-pair libraries. A second *P. glauca* genome of a genotype used in breeding was also produced and compared with the aforementioned.

In contrast, the assembly of Norway spruce [Nystedt et al. 2013] combined fosmid pools

with shotgun sequencing methods of haploid and diploid genomes (from megagametophyte and leaf tissue respectively). Sequencing a haploid genome holds great advantages as there will be no polymorphic sites, and therefore in theory, no bubbles in the de bruijn graph (however repeat regions will still cause incorrect merges of assembled k-mers). A total of 148x coverage was generated for *P. abies*, also using a variety of insert sizes from 180 bp to 10 kbp, all sequenced using Illumina HiSeq. Some paired reads were also overlapped and merged in order to generate longer single reads. Paired reads were also sequenced from the fosmid pools and assembled separately. A hierarchical assembly strategy was employed to make best use of all read libraries; the fosmid assemblies were merged into the haploid assembly, and the resulting contigs were scaffolded using paired reads from the diploid tissue libraries. Heterozygosity from the diploid tissue was not incorporated into the genome assembly; this was likely chosen as a way to avoid the de bruijn graph from being overly complex, and producing a less contiguous assembly.

The assembly strategy of loblolly pine, *Pinus taeda*, is very similar to that of Norway spruce [Neale et al. 2014; Zimin et al. 2014]. Haploid DNA was extracted from the megagametophyte, and diploid DNA from needle tissue. Highly divergent haplotypes were filtered from the diploid data to reduce assembly complexity. Paired-end reads were assembled using an Overlap Layout Consensus assembler (see section 3.1.1), MaSuRCA [Zimin et al. 2013]. MaSuRCA allows a large reduction in the number of reads to be assembled, by condensing the haploid reads into super-reads that spanned the insert size of the original pairs. The computational power required to assemble the 'super-reads' was therefore feasible. The assembly was then scaffolded using independent genome and transcriptome assemblies in a second version. Very recently (August 2016), version 2.0 of the *P. taeda* genome was released as an early draft at pinegenome.org and on the Dendrome Project ftp server. It improves the N50 from 66 kbp to 107 kbp, but there are not yet detailed methods on how this was achieved.

2.2.3 Breeding for desirable traits - *Salix* spp.

(Adapted from Sollars & Buggs 2016).

Salix purpurea has become a key model species in genetic improvement for shrub willows, for use as a biomass crop [Fabio et al. 2016]. Shrub willow has many characteristics ideal for biomass production; high yield, resprouting after coppice, and being easy to breed and therefore improve. Currently, willow yields between 8 and 12 Mg/ha/year of dry mass [Volk et al. 2016]. Traits such as yield are expected to increase by 20-40% with the development of new improved varieties [Volk et al. 2016], a process which could be greatly aided by genetic information. A genome project is led by Larry Smart's group at Cornell University (http://willow.cals.cornell.edu/genome/) and involves researchers from Oak Ridge National Laboratory and the J. Craig Venter Genome Institute. The 450 Mbp genome has been sequenced using the Illumina platform and assembled into scaffolds which have been annotated using RNA-seq data and anchored to a genetic map (Carlson et al. 2014). This has been released at http://phytozome.jgi.doe.gov/. The genome has already been used as a reference for genotyping by sequencing (GBS) of hundreds of further individuals, leading to the identification of genetic markers associated with growth and biomass yield (Carlson et al. 2016;

Gouker et al. 2016). Novel cultivars with increased biomass yields have already been generated from genetic improvement methods. For example, triploid hybrids have been found to give higher yield and improved biomass composition than diploids [Serapiglia et al. 2014].

Another species in the *Salix* genus has also been sequenced; the genome of *Salix suchowensis* was published in 2014 by research institutes in China, USA, and the UK (Dai et al. 2014). It consists of 304 Mbp of sequence in 103,144 scaffolds, on which 26,599 putative protein-coding genes were annotated. They also compare the genome to that of poplar (*Populus trichocarpa*), and investigate divergence, substitution rates, and whole genome duplications in the two species.

2.2.4 Disease resistance - *Castanea mollissima*

(Taken from Sollars & Buggs 2016)

The Chinese chestnut has received particular attention as a genomic resource because the species is resistant to chestnut blight, a disease caused by the pathogenic fungus *Cryphonectria parasitica* [Anagnostakis 1987]. This fungus has devastated American chestnuts, which are highly susceptible, since its introduction to the USA around 1904 [Anagnostakis 1987]. Considerable effort has gone into breeding American chestnut trees with resistance to the fungus either through hybridising American with Japanese or Chinese chestnuts, or by using transgenics to introduce resistance genes into the American chestnut genome [Hebard et al. 2014].

Genetic and physical maps of the *Castanea mollissima* (approx. 800 Mbp) genome were published in 2013 [Kubisiak et al. 2013; Fang et al. 2013]. A consortium led by John Carlson at Penn State University has assembled a genome sequence of *C. mollissima* using a combination of 454 and Illumina MiSeq reads, and BAC paired-end Sanger sequences [Carlson 2014], with scaffolds anchored into pseudochromosomes using the physical map. They also annotated the genome with over 36,000 gene models. The consortium has sequenced additional genotypes of Chinese and American chestnut to obtain variant data [Carlson 2014]. A research group based at Purdue University has resequenced 16 Chinese chestnuts and hybrids with variable blight resistance, in order to investigate variation at loci implicated in resistance [LaBonte and Woeste 2016].

2.2.5 Population biology and phylogeography - *Betula* spp.

(Taken from Sollars & Buggs 2016)

Dwarf birch is a small tree found in boreal scrub communities; one of the most northerly distributed woody angiosperms. Though of little economic importance, it is a keystone species to the ecology of the sub-arctic. A draft assembly of the relatively small 450 Mbp genome was published in 2013 using Illumina sequencing [Wang et al. 2013]. Though fragmented and preliminary, the assembly was a useful reference for the restriction amplified digest (RAD) sequencing of other individuals of *B. nana*, as well as *B. pendula* and tetraploid species *B. pubescens*[Wang et al. 2013]. Genomic markers have indicated signatures of hybridisation and introgression between the three species across the UK [Zohren et al. 2016; Wang et al. 2013]. However, the direction of introgression found has depended on

the type of markers used; microsatellites show a bi-directional introgression; whereas RAD markers indicate uni-directional introgression from both diploid species to the tetraploid. An improved *B. nana* assembly using SMRT sequencing (Pacific Biosciences, CA, USA) is underway. In addition, projects assembling the genomes of *B. pendula* by a team from the University of Helsinki, and of *B. platyphylla* (http://birch.genomics.cn) by a collarborative team lead by Hairong Wei at Michigan Technological University. Neither projects have been published yet, but the genomic data of *B. platyphylla* is available to download.

2.2.6 Tree genome databases

(Adapted from Sollars & Buggs 2016)

Several projects have attempted to collate genomic data for many tree species together into common databases, which act as a resource hub for researchers to access data and tools in one place. One consideration with having so many tree genome projects currently in progress, is that data can become hidden away in researchers' own websites. For outside researchers, both finding the data in the first place, and being updated on when new versions are available can be cumbersome when each specie's genome is stored in a different place, and sometimes the same species is being studied by different groups. Having a large collection of data available in the same repository enables studies such as genomic improvement, population biology, evolution, forest health and comparative genomics to be carried out with greater ease, since these tend to require, or can be vastly improved by, using data from several species.

The Fagaceae project (http://www.fagaceae.org/) was one of the earlier multi-species genome projects, focusing on the Fagaceae family (oaks and chestnuts). EST assemblies and SSR candidates were assembled for eight beech, oak and chestnut species, as well as an online BLAST tool and database. Ending in 2010, it aimed to produce data that would aid the breeding of a chestnut tree resistant to chestnut blight, a disease caused by the pathogenic fungus *Cryphonectria parasitica* (Anagnostakis 1987). The Chinese chestnut (*Castanea mollissima*) received particular attention within the project because the species is resistant to the fungus.

The Fagaceae project overlaps greatly with Hardwood Genomics, a project initiated to house genome and transcriptome data for a number of hardwood species (many listed in Table 2.2). Most of the Fagaceae data is also hosted here, as well as visualisation tools for the Chinese chestnut physical map and genome. Again, the aim of the project was to provide improved access to tree genomes in order to aid studies into forest health and tree improvement. Transcriptomes for some species under various stresses such as heat, drought, and disease, have been analysed (e.g., for green ash [Lane et al. 2016]), which could give insights into the expression-level responses of forest trees to changing climatic conditions.

Very recently, the researchers behind Hardwood Genomics were co-awarded a grant to develop a new database, Tripal Gateway, which will connect various plant genome sites already using the Tripal platform (NSF award #1443040). The new database will include the data already stored in Hardwood Genomics and TreeGenes (another database storing

genomic data for eighteen forest tree species, as well as transcriptome and proteome data for many more), in addition to other plant genome collections such as the Citrus Genome Database and CottonGen. As part of the project, Tripal will be integrated with bioinformatics software platform Galaxy, allowing common analytic tools and workflows to be executed on the data.

2.3 Conclusions

(Adapted from Sollars & Buggs 2016)

Herbaceous angiosperms have tended to dominate genomic research, as they are far more amenable to experimentation and breeding than trees. Nevertheless, rapid progress has been made in tree genomes over the past few years due to the cost reduction of next generation sequencing, and the pace of progress is set to increase as new technologies such as Oxford Nanopore (Oxford Nanopore Technologies, Oxford, UK) sequencing, SMRT (Single Molecule Real Time) sequencing (Pacific Biosciences, CA, USA), and optical mapping continue to improve [Howe and Wood 2015; VanBuren et al. 2015]. Current reference genome assemblies still need to be improved, as most of the genome sequences reviewed in Section 2.2 are still in fragmented states and far from being assembled at a chromosomal level. A better genome assembly may lead to more powerful genome-wide analyses of, for example, trait-associated loci or patterns of introgression. Species- and genus-wide sequencing of multiple individuals will also benefit from a contiguous reference sequence. However, current technologies are limiting genomes with high heterozygosity, repeat content, or polyploidy, from being sequenced and assembled contiguously. Future enhancements of sufficient magnitude are likely to stem from technologies that focus on joining and ordering scaffolds, such as optical mapping and BioNano Irys, or filling in assembly gaps that the usual technologies are unable to sequence, rather than from simply obtaining additional sequencing data. Therefore, researchers should perhaps wait for these new technologies to become available or validated, before investing repeatedly in additional short read coverage.

There is also a need for research to focus on the functional characterisation of genes within the reference genome using experimental approaches. However, such approaches, which have worked well for *Arabidopsis* and other herbaceous species, are highly challenging in trees. The generation of inbred lines, knock-out lines, or multiple mapping populations are seldom feasible for long-lived trees that take many years to reach maturity. Building a knowledge base of tree-related gene functions would greatly benefit genomic selection and breeding. Genomic selection itself could save a great deal of time in experimental breeding by allowing selection to take place many years before a tree can be tested for a certain phenotype. In addition, newly developed, targeted methods could achieve what would take years with conventional selection approaches. For example, targeted mutagenesis by CRISPR/Cas9 has been used to create knockout mutations in *Populus tomentosa* (Chinese white poplar) [Fan et al. 2015], and the expression of genes was inhibited using virus-induced gene silencing in two other *Populus species* [Shen et al. 2015].

As well as holding great promise, genomic research on a wide range of trees is necessary and timely [Neale and Kremer 2011]. Forest trees can now be viewed as potential

crop species, as the resources and methods that enable selection and breeding for timber-associated traits are under development. The value of natural capital and the need for renewable energy sources and carbon fixation is now widely acknowledged [Helm 2015]. With land-use discrepancies threatening the existence of many large forests, research into tree ecology and genetics will also aid the conservation and understanding of the forest ecosystem as a whole.

Chapter 3

De novo genome assembly and annotation of a British *Fraxinus excelsior* tree

3.1 Introduction to genome assembly and finishing methods

3.1.1 de Bruijn graphs vs Overlap Layout Consensus methods

Before short-read NGS technologies became as ubiquitous as they are today, most new genomes were sequenced using Sanger sequencing. Although the process was more time-consuming than current technologies, the Sanger reads were longer and there were fewer of them. Genome assembling software for these types of reads focused on the Overlap Layout Consensus (OLC) algorithm. This method performed a pairwise comparison of every read against every other, and gradually built a sequence from those that overlapped. With the advent of NGS, the computational and time requirements of an all-against-all search became too large. The all-against-all process scales quadratically, $O(N^2)$ with the number of sequences analysed (i.e., the computational time required is the square of the number of sequences).

De bruijn graphs were soon implemented into assembly software handle with the vast numbers of reads generated by NGS platforms. De bruijn graphs are directed graphs that describe the overlap between sequences. Instead of comparing the reads directly, they first partition the reads into k-mers (sub-sequences of length k). The distribution of k-mer frequency across all reads should show two peaks for a diploid genome; one for the homozygous sequence content around the frequency of the average genome coverage, and one for the heterozygous genome content at approximately half the frequency of the homozygous peak.

De bruijn graphs assemble genomes from the pool of k-mers. An overview of a representative pipeline, the CLC *de novo* assembly tool, is shown in Fig. 3.1. Firstly a word table is produced using all k-mers. Using one k-mer as a starting point, all possible backward and forward neighbours are generated (Fig. 3.1A). In most cases, only one backward and one forward neighbour will be found. The de bruijn graph can be established, where the k-mers make up the graph nodes and edges connect neighbouring k-mers. Over successive k-mer

overlaps the graph can be extended, and nodes reduced where the sequence is linear and does not encounter polymorphisms or repeats.

Figure 3.1: De Bruijn graph assembly process. A). Word table showing the starting k-mer, with all possible forward and backward neighbours. B). The de bruijn graph splits when a SNP is encountered. Each path of the bubble contains one of the alleles, i.e. top path contains the C base, and the other contains the G. C). Two separate graphs are merged when a repeat region is found, as they share the same k-mer for that region. Images modified from CLC Assembly Cell White Paper, June 2016 (http://www.clcbio.com/support/white-papers/).

When a polymorphism or sequencing error occurs in the sequence, the graph then finds two potential neighbours for a k-mer. A bubble is created with two paths, one for each allele (Fig. 3.1B). In many cases, bubbles can be collapsed by choosing one of the paths through the graph, usually that with the most read support. In this way, sequencing errors are excluded, as these will have very low read support compared to heterozygous sites which will have roughly equal read support for each allele. Highly heterozygous genomes result in a great number of bubbles and of very large sizes, making them more difficult to collapse into a single contiguous sequence. If the bubble cannot be collapsed (for example if there is a long stretch of heterozygosity), then the graph will be split and the two alleles will be two separate contigs in the resulting assembly. If this happens in many regions, the assembly can become larger and will contain many small duplicated contigs.

In contrast, repeat regions cause a merging of the graph from two different places that both share the same neighbour (Fig. 3.1C). These can also usually be resolved, if there are reads long enough to span the repeat region, or by using long-range mate pairs. Similarly to a heterozygous bubble, if a merge in the graph cannot be resolved, the de bruijn graph is split into separate contigs which reduces the contiguity of the assembly, and can reduce the assembly size (because the repeat region is only present in one copy). Assemblies of

repetitive or heterozygous genomes are therefore often very fragmented, with several small contigs of near identical sequences.

Increasing the k-mer size increases the number of possible distinct k-mers and can therefore cause the overall assembly size to expand. Longer k-mers are more likely to contain at least one locus that is polymorphic (or a sequencing error), decreasing the linearity of the de bruijn graph. Usually bubbles can be collapsed as explained previously. However, in very complex areas of the graph this may not be possible, such as in areas of high SNP density. Additionally, longer k-mers increase the chance that a k-mer will contain more than one polymorphic locus. If the SNPs are linked (i.e., on the same allele) then the bubble can be extended to include all SNPs with one path for each allele. Otherwise, bubbles can become branched in order to incorporate all possible paths linking the SNPs. Complex bubbles are more difficult to collapse into a single representative haplotype, and therefore the graph may become broken and the contigs split. Longer k-mers also increase the chances of spanning small repeats (those shorter than the length of k) and thus being able to assemble them along their entire length without being shortened by mistake.

Gaps in the assembly can arise due to many reasons. Library preparation methods often miss some regions in the genome, and sequencing is not always successful for every DNA fragment. Therefore, these regions will be missing in the raw DNA reads and will not be incorporated into the assembly. Repetitive regions (such as telomeres or short sequence repeats) are particularly difficult to sequence using short sequence reads and also make it difficult to resolve the de bruijn graph as k-mers become less complex further into the repeat. In the case of long-range repeats, such as paralogous regions resulting from genome duplication, it can be difficult to discern the correct path through the de bruijn graph when the same k-mers originate from different regions in the genome. Another possibility is that some parts of the genome were sequenced in isolation, i.e., the surrounding regions were not sequenced. Although the small 'island' of sequence could be assembled, it cannot be joined onto any other part of the assembly or could be too small to pass the minimum contig length threshold. Finally, low coverage regions may be ignored from the contig assembly altogether, as these paths will have low read support in the de bruijn graph compared to paths with normal coverage and could be treated as erroneous.

Further genome analyses become more difficult with a fragmented, highly duplicated assembly. Read mapping results in many reads that can be placed in more than one location (non-unique reads). Similarly, if a repeat region is collapsed into a single sequence, reads from the different locations will map to the same sequence when they should be separated. Calling variants in either of these areas can lead to erroneous results, because the reads do not correspond to the correct part of the reference sequence. In cases where polymorphisms have caused the graph to split into separate contigs, fewer variants will be called, because the reads from each allele will map to their respective contigs near perfectly. On the other hand, a large number of variants may be called in a collapsed repeat region because the reads from other locations (which may not be polymorphic in that region) may be slightly different from the collapsed sequence. However, these can sometimes be identified by a large jump in the coverage. Paired reads with large insert sizes may become broken, if they map

to separate contigs. This can often happen when large numbers of contigs are smaller than the insert size. Usually the reads can still be mapped as singletons, however the placing is sometimes not as accurate as when the pairing information is used.

3.1.2 Scaffolding and gap filling

Once contigs are generated, the information held in paired reads can be used to join contigs together. The distance between the paired reads in each library is known (to a degree) or can be estimated from the mapping of reads within a contig. Therefore, in cases where members of a pair map to the edges of two different contigs, the two contigs can be joined with the expected number of 'N' nucleotides that would place the two reads the expected distance away from each other. If the estimated insert size is incorrect however, then the wrong number of Ns could be placed between contigs. Though usually a range of insert sizes is specified and if reads from multiple insert libraries span the gap then a consensus gap size can be calculated. Reads with an incorrect orientation caused by erroneous library prep can be identified as they map differently to the majority of other reads, and are usually ignored.

SSPACE [Boetzer et al. 2011] is one such tool that performs scaffolding. It first maps paired reads starting from the smallest insert library, to the assembled contigs using Bowtie [Langmead et al. 2009], and the information for location and orientation of pairs are stored. Contigs are then considered for pairing if there are a sufficient (user-defined) number of paired reads spanning the two contigs, and the distances between them fits with the expectation from the paired read insert sizes (also user-defined). Contigs are joined iteratively starting with the largest. The algorithm also tries to account for contigs with several possible pairings; it will either place the contigs in the order suggested from the read mappings and insert sizes (e.g., three contigs joined successively), or will choose the best pairing if one pair combination has more read support than the other (i.e., if one pair of contigs were paired erroneously). The whole process is then repeated hierarchically with reads of the next smallest insert size, and so on until no more contigs can be joined into scaffolds.

The gaps of 'N's between contigs can then be 'filled in' with sequence. Once again, the information in paired reads can be utilised here. In some cases, one member of a pair will map within the contig and the other would be expected to map within the gap region, based on the expected insert size. Using many sets of paired reads, a large amount of gap-mapping reads can be placed within the gap, and a specific *de novo* assembly can be carried out with these reads to fill at least some of the gap sequence. The assembled 'island' can then be extended using iterative read mapping. This is a method used by gap closing programs GapCloser from SOAPdenovo2 [Luo et al. 2012] and GapFiller program from BaseClear [Boetzer & Pirovano 2012].

Recent advances in long read technology, such as Pacific Biosciences, mean that often research groups obtain long-read data for their studied genome to combine with initial *de novo* assemblies, usually generated using short reads such as Illumina. PBJelly [English et al. 2012] is one such tool that performs gap filling using long reads. Initially designed for

long PacBio reads to finish small bacterial genomes, it can also handle 454 reads (which were available to us in the ash genome project) and can at least make steps towards finishing genomes that are much larger. PBjelly works by extending the edges of contigs into the gaps, using the read mapping to generate a consensus sequence to fill at least part of the gap if not all. If the gap is too large to be closed, PBjelly will output the extension of the flanking sequences to reduce the gap's size.

3.1.3 Assembly verification and comparison

Many assemblies can be generated within just a few days (depending on data size and compute power, multiple ash assemblies could be generated within a day) by tweaking parameters within a number of different assembly or genome finishing programs. Due to the nature of *de novo* assemblies, we cannot know which assembly is the most true to the real genome. In some cases, we can use a very closely related species that already has a genome sequence available, to provide a comparison tool for the set of generated assemblies. However, this is often not the case (for example, the closest species to ash with a genome sequence available at the time was monkey flower, which only shares membership of the family Lamiales) and other comparisons need to be made.

Simple length statistics are a good way to gain an indication of the contiguity of the assembly. Measures such as longest scaffold and total assembly size are a quick and easy way to compare between versions. If the total size of the assembly is far larger than the estimated genome size, the software has assembled too much sequence. This can be caused by high heterozygosity in the genome leading to the assembly of alleles as if they were paralogous, as explained in Section 3.1.1. The mean scaffold size is not a very good indication of contiguity, as the mean is skewed by a few very large scaffolds or by lots of small ones. Similarly, the median is always pulled downwards by a large number of small scaffolds. A measure used more often is the N50 - the contig size at which half of the assembled bases are present, when the contigs are ordered by size. In other words, the N50 is the size of the contig that contains the median base pair. However, using the size of the assembly as a means to calculate the N50 is inherently affected by the total size of the assembly. The NG50 is known to be a more useful measure; instead of using the total size of the *assembly*, it finds the median base pair of the estimated *genome* size. In this way, the median base pair is always the same absolute distance from the first base of the longest contig, and is not influenced by the size of the assembly.

Length statistics however, can only give an indication of the contiguity of the assembly rather than its accuracy. It can be easy to change assembly parameters to produce large N50s, but if the longer scaffolds are made of misassembled bases then the long scaffolds are of little use. A mapping of the original sequencing reads to the assembly can provide more information on the assembly accuracy. If a higher percentage of reads map to one version than the other and a higher number of paired reads remain intact, then the software has assembled these reads within the same scaffold and the expected distance apart. Other tools can detect misassemblies from read mappings. For example, FRCbam [Vezzi et al. 2012] uses

a Feature Response (FR) curve to show the accumulation of misassemblies over the length of the assembly. By plotting the number of 'features' against the cumulative scaffold length, the FR curves of many different assemblies can be visualised and compared. The assembly showing the steepest, most r-shaped curve, incorporates the fewest misassemblies over the assembly length. The features include measures such as low or high coverage regions, reads with unmapped pairs, broken pairs, or paired reads with the incorrect orientation.

One downside of these read mapping statistics is that they rely on the starting data, the sequenced reads, to generate the comparisons, and are therefore vulnerable to any biases that may be present in the read data. Other tools exist that do not depend on read data; CEGMA [Parra et al. 2007] compares assemblies using a core set of eukaryotic genes, the NCBI euKaryotic cluster of Orthologous Groups (KOGs). Being considered orthologous throughout the eukaryotic kingdom, these genes should be present in all eukaryotic genome assemblies. However, only six species were used to generate the 'core' set of eukaryotic genes, and only one plant. Therefore some of these genes may not be common to *all* eukaryotes. In CEGMA, the genes are tested against the assembly using a BLAST search, and reported as either complete (>70% protein length) or partial matches. Recently, support for CEGMA ended and the authors recommend instead using BUSCO [Simao et al. 2015]. BUSCO uses universal single-copy orthologs from OrthoDB, and can be implemented using lineage specific sets of orthologs. This feature enables a more specific assembly QC by allowing plant-specific orthologs to be tested. As BUSCO was only published in June 2015, and even now the plant dataset is only in early access phase, we did not use it on our ash assemblies. We did however use CEGMA to gain an indication of assembly completeness.

In this chapter, I present the process and results of assembling the *F. excelsior* genome from whole genome sequencing reads. Various versions of the assembly are detailed which all improved on the previous, by either incorporating new data or from optimising software parameters. I verify each assembly using length and read mappingstatistics, alignments to core eukaryotic genes using CEGMA, and in some cases, using FR curves. An initial draft gene annotation is also presented. Finally, I discuss potential technologies that could be used to improve the contiguity of the ash genome.

3.2 Methods

3.2.1 DNA Extraction and sequencing

A low heterozygosity hermaphroditic tree, '2451', from a wood in Oxfordshire, UK, was used in a controlled self-pollination by David Boshier (Oxford University) in March 2003. This tree had the lowest heterozygosity at four microsatellite loci in comparison to 18 other trees in the same woodland (D. Boshier, unpublished data). Seedlings from the self-pollination were established in February 2007 at The Earth Trust's Paradise Wood in Oxfordshire, UK. Twig samples were collected from one progeny, 2451S, in January 2013 by Richard Buggs.

DNA was extracted from wood tissue by Jasmin Zohren at QMUL using CTAB and Qiagen DNeasy protocols. The genome size of this tree was measured at 877.24 ±1.41 Mbp

by flow cytometry. RNA was also extracted by Jasmin using the Qiagen RNeasy protocol from leaf tissue of tree 2451S and from leaf, cambium, root, and flower tissue of its parent tree in Gloucestershire. 2451S DNA was sequenced at Eurofins, Ebersberg, Germany in March 2013, using a mixed approached of two technologies: 1) low coverage 454 GS FLX+ pyrosequencing, mean read length of 642 bp, and 2) high coverage Illumina HiSeq 2000 with a read length of 100 bp and a mixture of various short insert libraries and Long Jumping Distance (LJD) libraries. LJD libraries are created from extracting two distant loci on a chromosome and deleting some of the sequence inbetween. The two sequences are then joined and circularised, and the original ends can be sequenced in a forward-reverse manner. An additional Nextera mate-paired library was sequenced in February 2015 on an Illumina MiSeq at The Genome Analysis Centre (TGAC), UK. Mate-pair libraries differ from LJD libraries because they do not span such long distances as LJD, primarily because no sequence is removed from the two end loci, and the resulting paired reads are in a reverse-forward direction. All DNA reads were deposited into the European Nucleotide Archive with project code PRJEB4958.

All reads were quality trimmed and filtered using CLC Genomics Workbench (versions 6-8 depending on when data were received, though the 'Trim Sequences' tool did not change during this time) using the following parameters: minimum quality threshold 0.01 (equivalent to Phred score of 20), maximum of one 'N' nucleotide, and minimum length of 50 bp. Adapter and telomere sequences ('AAACCCT' repeats) were also trimmed off reads.

3.2.2 *De novo* assembly

The first version (BATG-0.1) was a rapid release after receiving the first installment of data from Eurofins; this consisted of only the low-coverage, long-read 454 data. The reads were assembled using 454 GSassembler v2.7 (http://www.454.com/products/analysis-software/) with the following parameters: '-sl 32 -urt -m -e 5'. The second version (BATG-0.2) was also a quick release after receiving the Illumina HiSeq data, but contiguity was still significantly improved from the first version. The CLCbio assembler (CLC Genomics Workbench v6.0) was used with word size (k-mer) of 64, with the short insert libraries used to construct the de bruijn graph, and other LJD libraries and 454 contigs used as 'guidance only reads'. Scaffolding was also performed using CLC. GapCloser v1.12 from the SOAP package was used to fill gaps within scaffolds. BATG-0.3 used the CLC bio assembler (CLC Genomics Workbench v6.5) only to generate contigs (word size 50), using the LJD pairs and BATG-0.1 contigs as 'guidance only reads'. The open-source tool SSPACE Basic version 2.0 [Boetzer et al. 2011] was used for scaffolding with the LJD reads, and gaps filled using GapCloser v1.12. The fourth assembly version, BATG-0.4, assembled contigs in CLC (CLC Genomics Workbench v6.5) similarly to BATG-0.3 but using a bubble size of 5000 in order to collapse more bubbles. SSPACE Basic v2.0 was used with the parameter '-k 7' to scaffold contigs using all paired reads (default -k value is 10). This change allowed contigs to be scaffolded together with fewer reads joining them, and was therefore less strict whilst still retaining a good level of evidence for a join. The most recent assembly version, BATG-0.5, which was used for all further genome analyses, benefited from the addition of a MiSeq Nextera DNA library. Reads from the 200bp, 300bp, 500bp and Nextera libraries, as well as the 454 reads

(amplified *in silico* to increase coverage up to similar levels to the Illumina reads) were used to build the de bruijn graph in CLC Genomics Workbench v8.0. Amplification of the 454 contigs was necessary because the assembler uses read depth coverage to weight different paths through the graph. If a unique path is provided by the 454 contigs, but these only have a coverage of one, the path may be considered akin to a sequencing error and may be ignored, therefore a coverage similar to the Illumina reads is required. One downside to amplifying the 454 contigs *in silico* is that any assembly errors will also be amplified, but these may become corrected with the addition of the high coverage Illumina reads. All LJD library reads were used as 'guidance only reads' to generate contigs. SSPACE Basic v2.0 and GapCloser v1.12 were used again to scaffold and close gaps respectively. In addition, PBJelly v14.7.14 (https://sourceforge.net/projects/pb-jelly/) was used to fill more gaps using the long-read 454 data. The mitochondrial and plastid genomes were later separated from the main nuclear assembly, as described in Chapter 4. Assembly methods are summarised in Table 3.1.

Table 3.1: Methods for five versions of assembly. All other settings except for those mentioned were kept as default.

Assembly	Data & Software used
0.1	454 library assembled using GSassembler with parameters '-sl 32 -urt -m -e 5'.
0.2	Illumina HiSeq libraries and 454 library assembled using CLC assembler, word size of 64. CLC also used for scaffolding.
0.3	CLC assembler used to generate contigs using word size of 50, with only small insert libraries. LJD libaries and 454 (BATG-0.1) contigs used as 'guidance-only' reads. SSPACE used for scaffolding, and GapCloser for gap-filling with all paired reads.
0.4	CLC assembler used to generate contigs using word size of 50 and bubble size 5000, with only small insert libraries. LJD libraries and 454 reads (multiplied to make higher coverage) used as 'guidance-only' reads. SSPACE used for scaffolding with parameter '-k 7'. GapCloser used for gap-filling with all paired reads.
0.5	CLC assembler used to generate contigs using word size of 50 and bubble size 5000, with only small insert libraries and Nextera library. LJD libraries and 454 reads (multiplied to make higher coverage) used as 'guidance-only' reads. SSPACE used for scaffolding with parameters '-k 7'. GapCloser used for gap-filling with all paired reads, and PBjelly used with 454 reads. Mt and Cp genomes separated out, see Chapter 4

Although many assembly parameters were tested for all assembly versions, several assembly programs and uses of sequencing reads were tested more thoroughly in versions 0.3 and 0.5 than others. In BATG-0.3, I tested three different pieces of software; the CLC bio assembler (CLC Genomics Workbench v6.5), the stand-alone scaffolding tool SSPACE Basic v2.0 [Boetzer et al. 2011], and the SOAPdenovo2 pipeline [Luo et al. 2012]. After running many assemblies with different parameters for all three pieces of software, I compared the best assembly from each using simple assembly statistics and the program FRCbam [Vezzi et al. 2012].

Different software combinations were also tested for the BATG-0.5 release. This version was marked by the addition of the Illumina Nextera library reads which were sequenced on

an Illumina MiSeq at TGAC. Unfortunately a large number of these were filtered out during QC leaving an average of 3.5x genome coverage, from the raw 16.2x coverage. This could possibly have been due to failed library prep or low quality starting DNA, causing a great deal of primers or adapters to be sequenced, or 'N' nucleotides to be included, and then get filtered out in the QC stage. This library had been planned to help with scaffolding contigs together, and therefore I first tried scaffolding the contigs generated in BATG-0.4 using these reads in addition to the HiSeq data, with the tools SSPACE Basic v2.0 and GapCloser v1.12 (denoted as 0.5d in the results section). I additionally used PBjelly v14.7.14 to join these scaffolds together (0.5e). The Nextera reads were also used in the *de novo* assembly to generate contigs using CLC Genomics Workbench v8.0, as well as scaffolding with SSPACE Basic Version 2.0, and gap closing with GapCloser v1.12 (0.5a). Again, PBjelly v14.7.14 was used in addition to the other tools to join scaffolds together (0.5b). As all assemblies had so far been producing a large number of 'N' nucleotides due to filling gaps between contigs, I also tried scaffolding contigs without using the longest range (40 kbp) LJD library (0.5c).

3.2.3 Gene annotation

Although gene annotation is possible without the use of RNA data, (*de novo* methods), transcriptome data add a layer of support to the annotation of genes, helping to identify splice sites and different transcript isoforms, as well as erroneous annotations. RNA was extracted from five ash samples; four from the parent tree of 2451S (using root, cambium, flower and leaf tissue), and one from 2451S itself using leaf tissue. The extraction was performed by Jasmin Zohren at QMUL using the Qiagen RNeasy protocol. The five samples were sequenced paired-end on Illumina HiSeq 2000, using a 200 bp insert size.

Raw RNA reads were trimmed and filtered using CLC Genomics Workbench v6.5 with the following parameters: minimum quality of 0.01 (equivalent to Phred score of 20) and minimum length of 50 bp. Adapter and telomeric sequences were also removed. Filtered reads were mapped to the BATG-0.4 assembly (the most recently released version at the time), using the 'Large Gap Read Mapping' tool in CLC Genomics Workbench v6.5, which accounts for intronic gaps in the mRNA reads. Transcripts were then predicted using the 'Transcript Discovery' tool, and one (the longest) was selected as the single reference 'unigene' sequence for each gene, if multiple transcripts existed. RNA reads were then mapped back to the unigene transcripts and those with an average coverage along their whole length of less than 5x were filtered out.

3.3 *De novo* assembly results

3.3.1 Overall comparison of released assemblies

Sequencing yields and approximate genome coverage of the nine HiSeq, MiSeq and 454 libraries are shown in Table 3.2.

Table 3.2: Sequencing yield of 2451S using three technologies; 454, HiSeq and MiSeq. Genome coverage column describes approximate raw coverage of the 880 Mbp genome.

Technology	Insert Size	Read Length	Millions raw reads	Genome Coverage
454	n/a	Mean 642 bp	6	4.5x
HiSeq	200 bp	100 bp	350	40x
HiSeq	300 bp	100 bp	138	15x
HiSeq	500 bp	100 bp	245	30x
HiSeq	3 kb	100 bp	147	15x
HiSeq	8 kb LJD	100 bp	328	35x
HiSeq	20 kb LJD	100 bp	348	40x
HiSeq	40 kbp LJD	100 bp	256	30x
MiSeq	5 kbp Nextera	300 bp	47	16.2x

Five versions of the ash genome assembly have so far been released on our public and open website, ashgenome.org. Each successive version has improved assembly statistics by either incorporating new data or by tweaking certain parameters in the assembly software. An overall comparison of the five assembly statistics is shown in Table 3.3.

Table 3.3: Comparison of five ash genome assemblies

Statistic	BATG-0.1	BATG-0.2	BATG-0.3	BATG-0.4	BATG-0.5
Released	22/04/2013	11/06/2013	23/09/2013	11/11/2013	29/10/2015
No. of contigs	417,760	283,188	142,021	89,285	89,514
Total size	618 Mbp	1,469 Mbp	982 Mbp	875 Mbp	868 mbp
Longest contig	51,710 bp	221,212 bp	560,578 bp	696,341 bp	884,900 bp
Shortest contig	100 bp	344 bp	500 bp	500 bp	326 bp
No. >1 kbp	209,375	163,534	60,768	39,713	40,777
No. >10 kbp	1549	45,273	14,113	10,818	10,151
Mean size	1,480 bp	5,189 bp	6,917 bp	9,803 bp	9,691
Median size	1,004 bp	1,266 bp	866 bp	878 bp	911
N50 length	2,412 bp	14,228 bp	68,494 bp	98,766 bp	103,995 bp
L50 count	73,440	30,458	3,996	2,526	2,389
%A	32.6	27.16	24.07	26.54	27.22
%C	17.38	14.17	12.54	13.83	14.19
%G	17.15	14.16	12.54	13.84	14.19
%T	32.86	27.17	24.05	26.52	27.20
%N	0	17.34	26.81	19.27	17.19
CEGMA complete	127 (51%)	189 (76%)	214 (86%)	220 (89%)	208 (84%)
CEGMA partial	220 (89%)	237 (96%)	242 (98%)	241 (97%)	238 (96%)

3.3.2 Testing different software

Different software combinations were tested more thoroughly for assembly versions 0.3 and 0.5. Results for version 0.3 comparisons are shown in Table 3.4 and Fig. 3.2. The figure also incorporates the BATG-0.3 assembly that was released on the ash genome website. This released version is an improvement on the CLC+SSPACE version used during testing, after parameters continued to be optimised for this software combination.

Table 3.4 shows that the CLC+SSPACE assembly had the closest assembly size to the known genome size of 880 Mbp, the smallest number of scaffolds, longest scaffold, fewest

broken pairs, and highest CEGMA scores. CLC+SSPACE comes second only for the metrics NG50 and % reads mapped. Its FRcurve was neither the steepest nor shallowest, but it incorporated the fewest errors in total. Although the CLC assembly had the highest NG50, which would suggest it had the highest contiguity, the fact that the assembly is so much larger than the known genome size suggests a great number of duplicated contigs or under-assembled heterozygosity in the assembly. These can occur when a polymorphism causes a bubble in the de bruijn graph. When a preferred path through the graph cannot be ascertained, the bubble is split and almost identical copies of the sequence are kept. Considering the FRcurve in Fig 3.2, the CLC assembly has the fewest misassemblies compared to the other two at comparable genome coverage, as the curve is steeper. However, it also accumulates a larger number of misassembly features in total compared to the CLC+SSPACE assemblies. The SOAPdenovo2 assembly performed poorly in all of the metrics, and also had the shallowest FRcurve, meaning that it incorporated more errors per length of sequence. Therefore, the CLC+SSPACE combination was chosen as the best software pipeline to use for contig assembly and scaffolding, and the parameters were tested further. The released version of BATG-0.3 was a later version of this assembly, improved by tweaking parameters and the use of certain DNA libraries.

Table 3.4: Statistics of assemblies for testing version 0.3, using CLC bio assembler (CLC Genomics Workbench v6.5), CLC bio in conjunction with SSPACE Basic v2.0, and the SOAPdenovo 2 pipeline. NG50 is a more comparable version of the N50 metric, where the length of the contig that reaches half of the known genome size (in this case 880 Mbp) is used, instead of half the assembly size.

Statistic	CLC	CLC+SSPACE	SOAPdenovo2
NG50	93,223	78,383	45,525
Size (Mbp)	2,027	979	1,470
No. scaffolds	203,780	142,371	208,477
Longest scaffold	412,238	558,761	379,077
% reads mapped	94.54	92.69	73.49
% broken pairs	43.74	35.91	44.54
% CEGMA partial hits	96.37	97.18	90.73
% CEGMA complete hits	82.66	86.69	72.18

Version 0.5 comparisons are shown in Table 3.5 (shown at end of chapter) and the results of FRCbam for these five assembles are shown Fig. 3.3. The table shows that only 0.5c (the assembly that did not use the 40 kb LJD library in scaffolding) is quite different to the rest. It has a much lower %N than the others (8.89% compared to 17-18%), as expected. However, it also suffers from much lower contiguity than the other four, as shown by the low N50 and NG50 values (62 and 52 kbp respectively), shorter longest scaffold (578 kbp compared to >800kbp), higher number of scaffolds (96k compared to 86-89k) and lower total assembly size (774 Mbp compared to around 860 Mbp). As shown in Fig. 3.3, the FR curve for this assembly is the only to be separated from the rest of the group. This means that for the same genome coverage, assembly 0.5c incorporated more errors than the others.

It is slightly more difficult to choose the best assembly out of the remaining four, 0.5a, 0.5b, 0.5d and 0.5e, as each measured highest in different assembly statistics, and all FR curves are near identical. In addition, each pair of method-related assemblies (e.g. 0.5a & 0.5b, and 0.5d & 0.5e) show very similar statistics due to using almost the same data

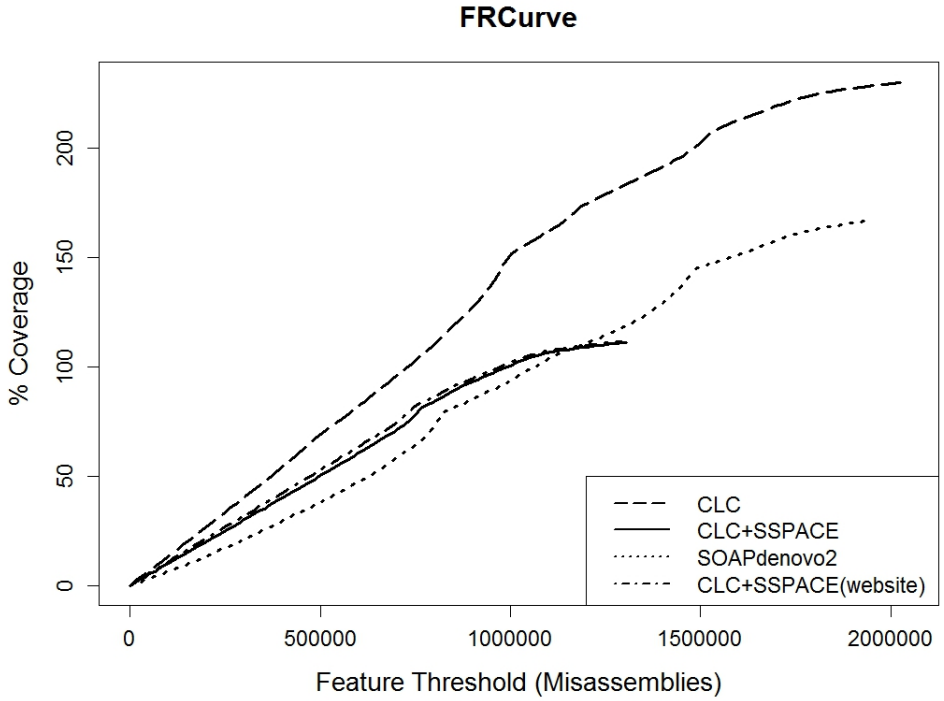


Figure 3.2: FR curve of software comparison for BATG-0.3, with lines representing the CLC bio assembler (CLC Genomics Workbench v6.5) used alone ("CLC"), CLC used with scaffolding tool SSPACE Basic v2.0 in testing ("CLC+SSPACE") and BATG-0.3 released version ("CLC+SSPACE(website)"), and the SOAPdenovo2 pipeline. FR curves depict the accumulation of misassemblies (features) over the length of the assembly. The assembly showing the steepest, most r-shaped curve, incorporates the fewest misassemblies over the assembly length. The features include measures such as low or high coverage regions, reads with unmapped pairs, broken pairs, or paired reads with the incorrect orientation. The released version of BATG-0.3 shows slight improvement over that during testing, as the curve is a little higher.

combinations. After much consideration, we (all QMUL and TGAC collaborators) chose 0.5b to be the 'best' assembly, and this version was released as BATG-0.5 after further improvements including assembling the organellar scaffolds (detailed in Chapter 4). We chose this on the basis of lower %N (i.e. more genomic sequence would be present) and higher percentages of reads mapping. All other statistics for this assembly, though perhaps not the best of the group, were sufficiently good to still be a large improvement from BATG-0.4; primarily, achieving an N50 of >100 kbp.

3.4 RNA-seq aided annotation of genes

In total, 36,944 transcripts were predicted, with a further break down for each sample shown in Table 3.6. These mRNA assemblies were released on the ash genome website, ashgenome.org, as well as proteins predicted using OrfPredictor [Min et al. 2005], a GFF3 annotation file, and unigene sequences (the longest transcript per gene to act as a reference transcript).

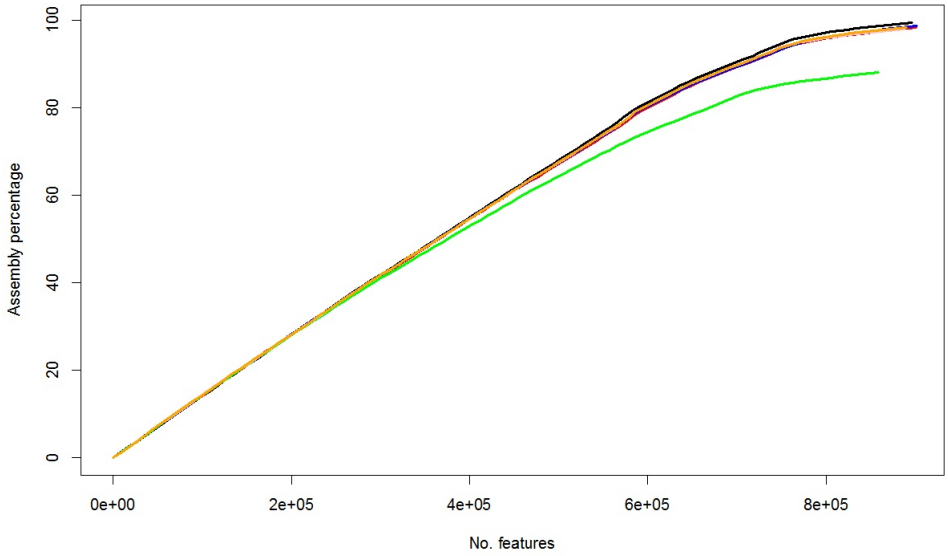


Figure 3.3: FR curve of software comparisons for BATG-0.5, with green line representing assembly 0.5c (without using 40kb LJD library for scaffolding), and all other assemblies indistinguishable in the higher lines.

Table 3.6: RNA sequencing yield of five ash samples: four from parent tree (root, cambium, leaf and flower) and one from 2451S (leaf tissue)

Sample	Tissue	No. raw reads	No. filtered reads	No. predicted genes
S_L1	2451S Leaf	66,466,032	62,731,876	27,360
M_L1	Parent Leaf	34,591,166	32,606,695	24,473
M_R2	Parent Root	80,587,054	76,199,721	28,275
M_C1	Parent Cambium	65,088,170	61,489,088	27,368
M_F1	Parent Flower	74,709,866	70,615,880	29,562

The assembly was used by collaborators Andrea Harper and Ian Bancroft as a reference for mapping mRNA reads and predicting markers associated with ADB susceptibility [Harper et al. 2016]. In addition, the assembled RNA contigs were used as evidence in the next version of the genome annotation, performed by David Swarbreck and Gemy Kaithokottil at TGAC.

3.5 Conclusion

De novo assembly of genomes is a difficult task that is never completely finished; assemblies can be continuously refined and improved by adding more data or using different methods to sequence or visualise the genome. Even for model organisms with long-standing genome projects, such as the human genome and that of *Arabidopsis thaliana*, new versions are released every few years with various improvements and rearrangements. The ash tree genome project has a great deal less funding and resources (e.g., existing genetic data) than that of model organisms, and despite best efforts, the genome is still in a draft, fragmented state. The five assemblies released represent continued work on improving this draft genome as-

sembly, along with efforts to make data as open and accessible as possible. Since the genome sequence has been released, others have been able to use it for further analyses, such as repeat content analysis and synteny analysis with other plant species (performed by Laura Kelly and reported in Sollars et al. *in press.*) In addition to the markers for ADB susceptibility already identified in Harper et al. (2016), others were found to be associated with susceptibility in a larger panel of trees in Sollars et al. (*in press*), in work performed by Andrea Harper and Ian Bancroft at the University of York. The genome sequence is currently being used in a genus-wide study of ADB and Emerald Ash Borer (EAB) susceptibility in 31 *Fraxinus* species, for which it will provide a reference sequence to aid assembly of the other species' genomes.

Despite its proven usefulness in other studies, the assembly is still very fragmented. It is likely that heterozygosity in the genome as well as repeat regions have lead to misassemblies in the sequence. These are difficult to iron out without the use of high quality sequence data such as Sanger sequencing, which would be expensive and time-consuming to perform on a large genome. However it could be worthwhile to design primers into the gaps between contigs to generate sequence for those regions. Nevertheless, nearly 37,000 transcripts were annotated on the assembled sequence and CEGMA scores were also fairly high. Therefore it seems that a large amount of gene content is present.

Future improvements could include the addition of long-read data. As discussed in Chapter 2, generating additional high coverage short reads such as Illumina does not produce sufficient assembly improvement when the same money could be invested in a different type of data. Available long-read technologies include Pacific Biosciences and Oxford Nanopore. Although these technologies have fairly high error rates (around 15% for PacBio), there has been much success in using Illumina reads to correct the long PacBio reads, and then assembling these corrected reads. Some bioinformatic pipelines have already been developed using this method, such as proovread [Hackl et al. 2014], LoRDEC [Salmela & Rivals 2014], and the CLC PacBio assembly pipeline.

Another technology that could aid assembly improvement is optical mapping, such as BioNano Irys. Optical mapping generates an image of restriction sites along the genome, along which the *de novo* assembly scaffolds can be ordered. Although these data are available for the ash genome, we found that our assembly was not quite contiguous enough to produce meaningful results (work performed by Endymion Cooper, QMUL). Similarly, scaffolds could be anchored to a genetic map. However, this resource is not currently available for ash as it would require developing a mapping population of trees to measure linkage.

In conclusion, there are numerous possible routes to follow in order to improve the ash tree assembly. However, these would require additional funding and/or would need a more contiguous assembly to start from, such as for the use of BioNano data. Despite this, meaningful analyses have been carried out in this thesis using the latest genome release and gene annotation. Furthermore, the release of our assemblies on the project's website ashgenome.org has allowed other researchers to download and use the reference sequence and annotations for their own studies.

Table 3.5: Statistics of assemblies for testing version 0.5 using combinations of CLC (CLC Genomics Workbench v8.0), SSPACE Basic v2.0, GapCloser v1.12 and PBjelly v14.7.14. Read mapping was tested for each using the 500bp short insert library, Nextera library and 3kb LJD library. % pairs refers to percentage of read pairs kept intact upon mapping.

Statistic	0.5a	0.5b	0.5c	0.5d	0.5e
Notes	Contigs assembled using Nextera. Scaffolded using all data.	Same as left, but with PBjelly v14.7.14 using 454 reads	Contigs assembled with Nextera, scaffolded without 40kb data	0.4 contigs rescaffolded with additional Nextera data	Same as left, but also with PBjelly v14.7.14 using 454 reads to join scaffolds
Scaffold results					
Number	89,796	89,492	95,917	86,639	86,346
Total size(Mb)	864.8	867.9	774.1	862.3	865.4
Longest	884,631	884,900	578,045	847,940	847,214
Shortest	500	500	500	500	500
Number >1K	38,605	40,756	43,754	37,335	39,476
Number >10K	10,132	10,144	13,025	10,038	10,058
Number >100K	2,522	2,524	1,548	2,544	2,545
Mean size	9,631	9,698	8,071	9,952	10,022
Median size	857	911	899	855	912
N50 length	104,433	104,186	62,548	108,702	108,207
L50 count	2,373	2,386	3,526	2,292	2,305
NG50 length	101,974	102,149	51,947	105,654	105,778
%A	27.13	27.22	29.95	26.87	26.95
%C	14.15	14.19	15.62	14.00	14.04
%G	14.15	14.20	15.61	14.00	14.06
%T	27.13	27.20	29.93	26.85	26.96
%N	17.44	17.19	8.89	18.27	18.00
500bp library mapped (%)	91.43	91.61	91.26	91.11	91.40
% 500bp library pairs	77.37	77.52	76.99	76.85	77.02
Nextera reads mapped (%)	92.10	92.22	91.97	91.53	91.72
% Nextera reads in pairs	61.84	61.85	59.34	56.54	56.58
3kb library mapped (%)	91.35	91.51	91.19	91.11	91.44
% 3kb library pairs	57.80	57.87	57.23	57.59	57.71
CEGMA partial hits	237 (96%)	238 (96%)	237 (96%)	239 (96%)	240 (97%)
CEGMA complete hits	206 (83%)	208 (84%)	205 (83%)	207 (83%)	207 (83%)
Contig results					
Number	120,027	118,947	121,647	118,424	117,229
Total size (Mb)	714.0	718.8	705.4	704.7	709.7
Longest	237,620	237,620	237,620	220,678	225,516
Shortest	500	500	500	500	500
Number >1K	67,036	68,618	67,787	67,199	68,678
Number >10K	18,974	18,864	19,074	19,299	19,200
Number >100K	194	213	148	134	148
Mean size	5,949	6,043	5,799	5,951	6,054
Median size	1,183	1,228	1,176	1,217	1,261
N50 length	24,856	25,357	23,864	23,596	24,075
L50 count	8,020	7,913	8,311	8,415	8,283
NG50 length	17,586	18,064	16,423	16,305	16,916
contig %A	32.86	32.87	32.87	32.88	32.86
contig %C	17.14	17.14	17.14	17.13	17.12
contig %G	17.13	17.14	17.14	17.13	17.14
contig %T	32.86	32.85	32.85	32.85	32.87
contig %N	0.00	0.00	0.00	0.00	0.00

Chapter 4

Assembly and annotation of organellar genomes from whole genome sequencing reads

It is advantageous to separate the organellar genomes away from the nuclear genome for several reasons. Firstly, to avoid mis-assemblies where nuclear regions may be joined to similar regions of the organellar genomes by assembly software. Secondly, to perform analyses on the organellar genomes separately from the rest of the assembly; for example cpDNA is inherited uni-parentally and can therefore tell different stories about population history and structure compared to nuclear DNA, e.g., Section 6.3.1. In addition, some traits are specific to organelle genomes. For example, many genes involved in respiration and photosynthesis are located on mitochondria and plastid chromosomes. Therefore it is beneficial to know the sequences of these genes or loci, for example to develop markers for marker-assisted selection. Lastly, knowing which scaffolds in a draft assembly belong to the organellar genomes allows a user to exclude them from certain analyses that need not consider these regions. For example, some processes such as variant calling become very slow or stall in regions of high coverage (where there are lots of mapped reads to analyze), and therefore finish much quicker when these regions are not included in the first place.

High read coverage is typical of organellar genomes because the organelles themselves are present in high numbers within a cell compared to the single nucleus. Upon sequencing the whole cell's DNA content, the organellar genomes will have much higher read depth than the nuclear genome. K-mer (unique sequences of length k) peaks generated from whole genome sequencing data can help to predict the approximate read coverage of the organellar genomes. By plotting the occurrence frequency of k-mers (number of times the k-mer is observed) in raw sequencing reads against the number of distinct k-mers found at each depth, peaks in the graph can be identified. K-mers that peak at a low occurrence frequency usually correspond to the heterozygous and homozygous content of the nuclear genome, but smaller peaks at much higher frequencies will likely correspond to the organellar genomes, due to the higher number of organelles in each cell compared to the nucleus.

Sequencing the organellar DNA alone would first require isolation of the organelles themselves, followed by DNA extraction and sequencing. These steps can be time-consuming,

add to project costs for equipment, reagents and additional sequencing, and can often get contaminated with nuclear DNA [Ahmed & Fu 2015]. Therefore, it can be easier to sequence total DNA from a tissue sample, perform a whole *de novo* assembly and then identify which scaffolds belong to the organellar genomes. However, the assembly process often merges similar regions together during assembly; therefore, one scaffold could contain a mixture of nuclear and organellar DNA. Furthermore, DNA from one organellar genome could be spread over several scaffolds. Trying to BLAST a *de novo* assembly against a known mitochondrial or plastid sequence produces largely inconclusive results. In a preliminary analysis, I performed a BLAST alignment of the *Mimulus guttatus* mitochondrial genome against our whole genome assembly and found that many small regions match in multiple places, and many large scaffolds have small matches distributed throughout their length rather than focused in one region. Even if scaffolds have longer or a larger number of matches, it is difficult to define a threshold to definitively categorise a scaffold as organellar.

Therefore, I found it easier to first extract the DNA reads that I was confident belonged to either the mitochondrial or plastid genomes (due to BLAST score, explained in methods section), and then *de novo* assemble these separately. This chapter will describe the method used for extracting organellar reads out of a shotgun sequencing pool of DNA reads, and the assembly and annotation method for the mitochondrial and plastid genomes.

4.1 Methods

4.1.1 Generating k-mer distributions

I extracted 50 bp k-mers from the 200, 300 and 500 bp read libraries of the British ash tree, using the software Jellyfish v2.1.1 [Marcais & Kingsford 2011]. The program produces a list of these k-mers as well as the number of times they are found in the reads, and how many different k-mers are found the same number of times. I then plotted the k-mer frequency against the number of distinct k-mers found at each frequency in R v2.15.2. K-mers contributing to the high-coverage peaks were then used in a BLAST search (using CLC Genomics Workbench v7.0) to confirm that the peaks were indeed organellar.

4.1.2 Extracting organellar reads

Most organellar k-mers are located at frequencies above approximately 600x, which is where the slope of the mitochondrial k-mers starts to rise (presented in results section, Fig. 4.3). We can assume that if we extract any k-mers above this frequency, they would be mostly organellar. Therefore, every k-mer over 600x coverage was extracted as FASTA sequences and used in a BLAST search against the NCBI non-redundant (nr) database with a filter to allow matches only to plant sequences. K-mers were then extracted based on whether their first hit contained a mitochondrion or plastid / chloroplast related description, and separated depending on which organelle was the best match. The advantage of BLASTing only the high coverage k-mers (as opposed to just BLASTing every k-mer to be on the safe side), is that the number of query sequences is greatly reduced (276,000 compared to over 800 million), so the process finished much faster and downstream filtering is also quicker.

The result of this was two lists of k-mers, one with unique 50 bp sequences that were highly likely to be from the mitochondrial genome (over 191,000 k-mers), and one with unique 50 bp sequences highly likely to be from the plastid genome (over 131,000 k-mers).

Next, the sequencing reads that give rise to these k-mers were obtained. The advantage of going back to the reads as opposed to simply assembling the k-mers, is that reads are longer (90 bp after trimming compared to 50 bp k-mers), and they provide much higher coverage which is used in the assembly process. Reads from the 200, 300 and 500 bp libraries were filtered against the k-mer sets, and were kept if the first and last 50 bp matched any k-mers from either the plastid or mitochondrial k-mers (reads were at most 90 bp long, so some of the sequence in the middle of the read will be queried twice). This provides extra assurance that the read originates from the organellar genome and minimises the chance of including reads that could be ambiguous. Certainly, some true organellar reads will be excluded by these thresholds, or by the previous step of selecting only k-mers above 600x. There may also be gaps in the k-mer pool if the ash organelles contain sequences that are very different to other plant organellar genomes (i.e., unique to ash), so that the k-mer matches exceeded the e-value threshold during the BLAST search. The threshold allows a certain number of mismatches between the query sequence and the reference database, but too many mismatches will result in no hits and the k-mer getting excluded from the organellar collection. This is unlikely in the plastid genome since the sequence is very conserved, however it may cause gaps particularly in the mitochondrial assembly. However other reads and k-mers a few bases upstream or downstream will likely still cover the region; an advantage of having high coverage. Also in the case of the mitochondrial genome, the next steps of the assembly method were used to fill in gaps between and within assembled scaffolds using the whole genome 454 reads. These reads were used to extend out from the ends of existing contigs iteratively (see section 4.1.4), thereby incorporating any ash-specific sequence that did not pass the BLAST search. The iterative nature of the method allows the incorporation of unique ash sequence without any restriction on a maximum length of gap, as long as reads still map to the ends of the contigs, at least until the sequence extends enough to join up with another contig if this can be achieved.

4.1.3 Plastid genome assembly and annotation

Using the list of reads that matched plastid k-mers, a *de novo* assembly was performed using the CLC de novo assembly tool (CLC Genomics Workbench v7.0), with word size of 50 and bubble size of 5,000. The assembled contigs were aligned by BLAST to the olive (*Olea europaea*) plastid genome (RefSeq ID: NC_013707.2), and were able to be joined into one circular chromosome. Reads from the 454 library were mapped to the chromosome to check that read coverage was smooth across the sequence and to identify potential misassemblies. No large drops in read coverage or unaligned / overhanging ends were found, indicating that the assembly was largely correct according to the read mapping. One gap of approximately 30 N nucleotides was spanned by many 454 reads and the sequence could therefore be filled in manually.

With a few exceptions, all known plastid genomes possess an inverted repeat section (one

region that is repeated in the opposite direction further down the chromosome) of usually 25 kbp in length [Shaw et al. 2007]. These repeated sections split the remaining chromosome into the short single-copy region, and long single-copy region. Aligning the ash plastid assembly to the olive plastid also helped in arranging these repeat regions, as they would be expected to collapse into a single contig, following the de bruijn graph algorithm described in Section 3.1.1.

Genes were manually annotated on the sequence. A BLAST search was performed using gene sequences from two closely related species: *Olea europaea* (in the Oleaceae family), *Nicotiana tabacum* (in the Asterid clade). To my knowledge, these are the closet species to ash phylogenetically that have complete annotated chloroplast genomes, and therefore, should have the most conserved sequences to ash. In addition, gene sequences from model plant *Arabidopsis thaliana* were used, as this assembly would likely be the most studied and refined than other plants. The chloroplast chromosome was used in a BLAST search against the original *de novo* reference assembly to find out if the original assembler had created a separate plastid scaffold, and to replace it if so.

4.1.4 Mitochondrial genome assembly and annotation

Reads that matched mitochondrial k-mers (described in Section 4.1.2) were assembled using the CLC *de novo* assembly tool (CLC Genomics Workbench v7.0) with a word size of 50 and bubble size of 5,000. The initial assembly was very fragmented, consisting of 296 contigs with an N50 of 1.8 kbp. I wanted to improve the assembly by joining together contigs that might overlap or by filling in gaps between and within contigs. It is likely that some mitochondrial reads were missed in the extraction process due to the thresholds imposed earlier in the pipeline. Therefore the sequence information in the assembly gaps could still be obtained from the original pool of whole genome sequencing reads.

I put together a method to fill in gaps and join contigs together, using several tools in the CLC Genomics Workbench and the Genome Finishing Module plugin. First, 454 reads were mapped to the *de novo* assembled contigs. Contigs were then extended out from the edges using the overhanging reads, using the 'Extend Contigs' tool. Any overlapping contigs were joined using the 'Join Contigs' tool. A workflow and diagrammatic view of the method are shown in Fig. 4.1. Reads from the 454 library were mapped back to the new assembly to check that any joins made show seamless coverage (Fig 4.2); this ensures that joined contigs are truly adjacent as reads should carry on straight through the join. Gaps within contigs were also filled in manually using the overhangs of reads into the gap. Upon the next read mapping, the extended sequence was checked and extended further if possible (by extending the sequence, new reads map to the edges and overhang further into the gap). The overall map-extend-join method was performed iteratively, around ten times in total, with a significant improvement in assembly contiguity and reduction in gaps each time.

The advantages of using 454 over HiSeq reads are that 454 reads are much longer, thus being more likely to span a gap within and between contigs. They also have long overhangs from the edges of contigs which allows rapid extension out from contig ends on each iteration.

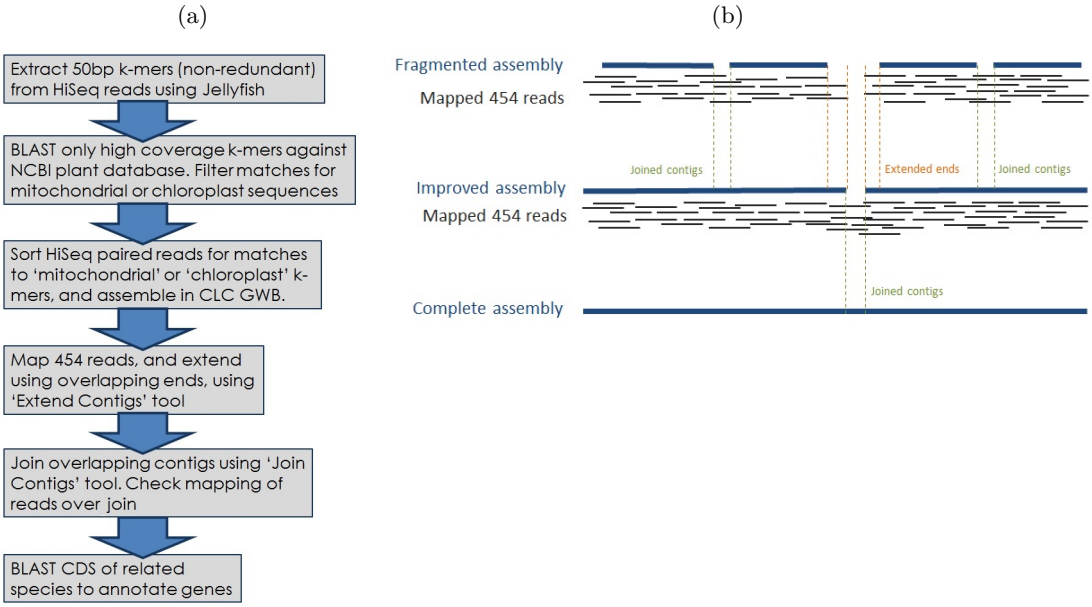


Figure 4.1: (a) Workflow of the map-extend-join method for mitochondrial genome assembly. (b) Mapped 454 reads are used to join contigs together and extend regions from the edges. Over several iterations, contiguity is increased until, in theory, a complete chromosome is produced.

Figure 4.2: Mapped reads can be used to check a join between contigs. If the coverage is smooth across the overlap region, the join can be trusted. Whereas if the reads only align to one side of the overlap region without crossing over into the next contig, this is likely a case of an incorrect join and the contigs should be split again.

4.2 Results

4.2.1 K-mer distributions reveal peaks of organellar sequence coverage

Two organellar peaks can be detected in plots of k-mer frequency (Fig 4.3 and 4.4). Figure 4.3 shows a peak of k-mer frequency at approximately 700x. K-mers around this frequency were BLASTed against the NCBI plant database, of which the vast majority matched mitochondrial sequences. Therefore, we can assume that this peak relates to the mitochondrial genome, and that its average coverage is around 700x.

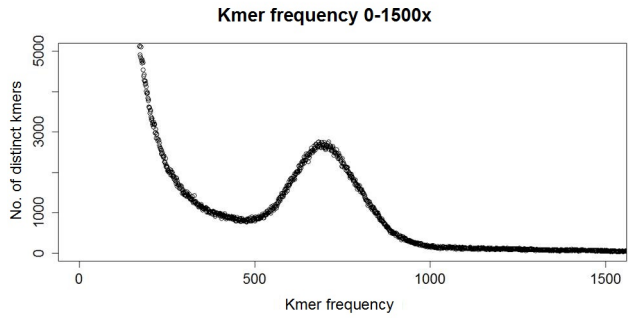


Figure 4.3: K-mer frequency between 0 (off-scale) and 1500, with a peak at around 600x. The vast majority of k-mers in this peak have top BLAST hits to other plant mitochondrial genomes.

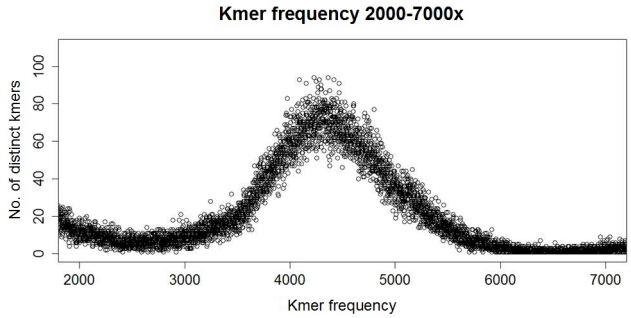


Figure 4.4: K-mer frequency between 2000 and 7000, with a peak at around 4200x. The vast majority of k-mers in this peak have top BLAST hits to other plastid genomes.

Figure 4.4 shows a peak of k-mer frequency at approximately 4200x. Results of the same BLAST search for these k-mers produced matches to mainly plastid sequences. Therefore we can assume that this peak relates to the plastid genome, and that its average coverage is around 4200x. This peak is at much higher k-mer frequency compared to the mitochondrial peak, which means that many more plastids were sequenced than mitochondria. The number of k-mers present in the plastid peak is also much lower compared to the mitochondrial peak (80 compared to 3000 at their highest). This means that the plastid genome is much smaller than the mitochondrial genome, as is commonplace in plants.

4.2.2 Assembly and annotation of the plastid genome

Two large contigs were assembled, one of 126,429 bp and one of 2,864 bp. In addition, a few more contigs were produced between 400 and 600 bp. To see if any contigs could be joined, the assembled contigs were used as a BLAST query against the *Olea europaea* chloroplast chromosome. *Olea europaea* is the closest plant species to ash phylogenetically that has an available chloroplast sequence, and as the chloroplast typically contains many conserved regions, the sequence of ash should be fairly similar. However it is not known how this sequence was assembled; whether it was sequenced in isolation or assembled from whole genome sequence reads. Nevertheless, it is still a useful reference to aid the ordering of the ash plastid contigs.

The two largest ash plastid contigs aligned very well to the olive plastid chromosome, and also shed light on the order and structure of the contigs. Fig. 4.5 shows that the largest contig appears split into five different pieces, however this is not actually the case. The first and last pieces were continuous in the ash sequence and were broken to show the alignment

in linear format against the olive plastid (i.e., the start and end of the olive plastid sequence, even though the plastid chromosome is circular). Another two pieces were the ends of the ash plastid sequence (as the default output of a *de novo* assembly is linear contigs) and these joined together when the sequence was circularised later. The orientation of one piece was the reverse of all the others, reflecting one area of the inverted repeat. Contig 2 aligned between some of the Contig 1 sequence. With some manual editing, the complete circular structure and sequence of the ash plastid sequence could be obtained (Fig. 4.6).

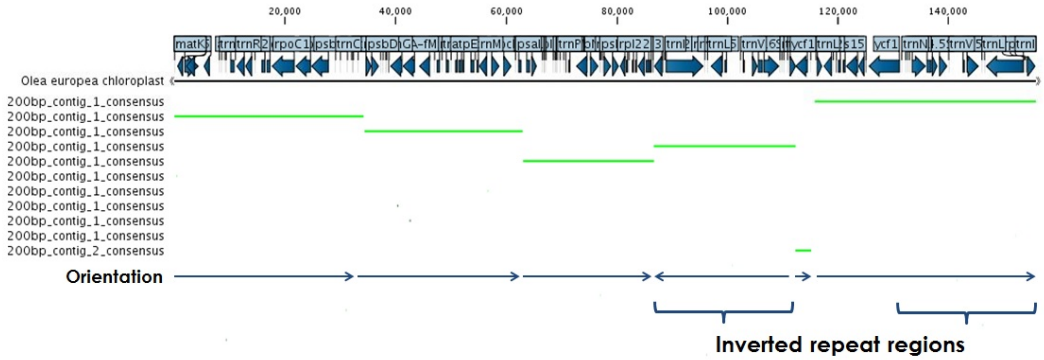


Figure 4.5: Alignment of assembled plastid contigs against olive plastid sequence. Contig 1 was split into five different pieces, although the first and last pieces were linear in the ash sequence and were broken to show the alignment in linear format. Another two pieces were the ends of the ash plastid sequence and these joined together when the sequence was circularised. The orientation of one piece was the reverse of all the others, reflecting one area of the inverted repeat. Contig 2 aligned between some of the Contig 1 sequence.

Using the alignments of *Olea europaea* and *Arabidopsis thaliana* plastid genes to guide manual annotations, 72 conserved protein-coding genes were annotated, as well as 16S, 23S and 5S rRNA genes, tRNA genes and 7 putative coding (ycf) genes. The full structural annotation of the plastid chromosome is shown in Fig. 4.6. Many plastid genes are of course related to the photosynthesis pathway e.g., NADH dehydrogenase (*ndh* genes), ATP synthase (*atp* genes), and RuBisCo large subunit (*rbcL*). RNA polymerase subunits (*rpo* genes) and many ribosomal proteins (*rpl* and *rps* genes) are also present. Four genes are located within the inverted repeat regions (*rps7*, *ndhB*, *rpl23* and *rpl2*) and are therefore present in two copies overall.

The chloroplast chromosome was used in a BLAST search against the original *de novo* reference assembly to find out if the original assembler had created a separate plastid scaffold. Indeed, a few scaffolds matched very well; Contig971 and Contig972 were removed from the *de novo* assembly, and regions of Contig4496, Contig970, Contig52351, Contig10012, Contig8882, Contig73223, Contig4712, Contig197, and Contig676 were deleted and replaced with the annotated chloroplast sequence.

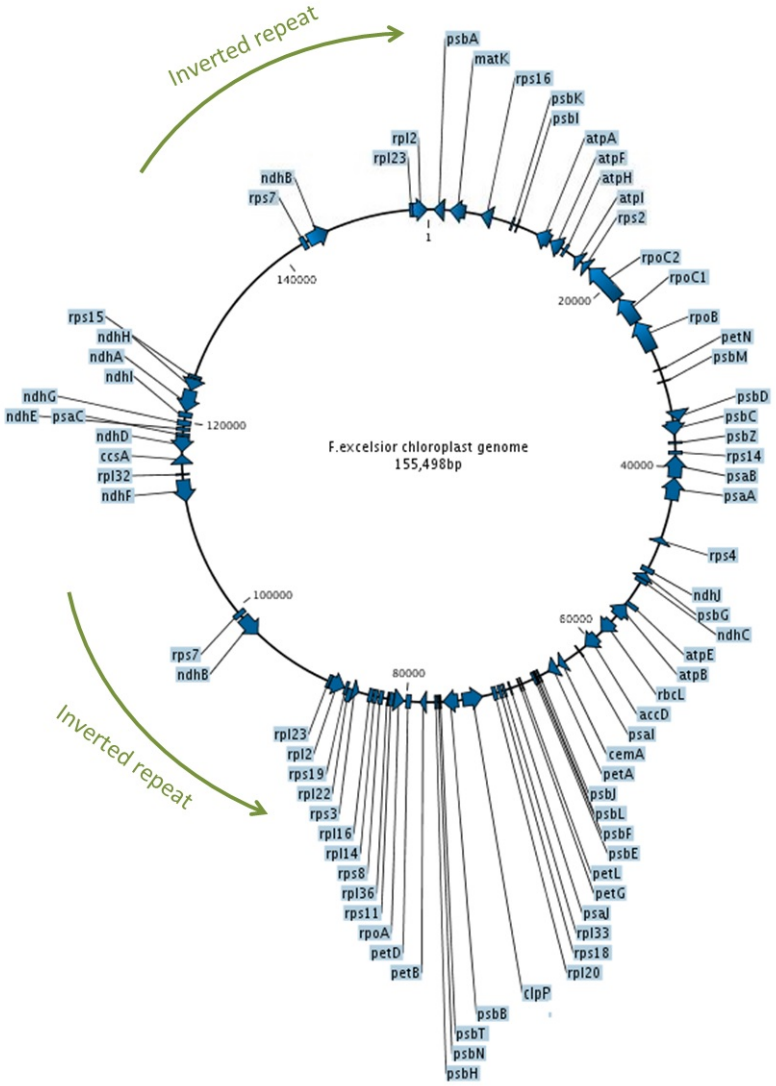


Figure 4.6: Gene annotation and structure of the plastid chromosome

4.2.3 Assembly of the mitochondrial genome using a map, extend and join method

After approximately ten iterations of the map-extend-join method, the assembly was much more contiguous, having been improved from its initial 296 contigs to only 26, with an N50 of 60 kbp (Table 4.1).

To ensure that the assembled sequence was truly mitochondrial, I attempted to annotate genes on the contigs. In a similar way to the plastid annotation, I used genes from related plant species *Mimulus guttatus* and *Nicotiana tabacum* as well as *Arabidopsis thaliana* in a BLAST search against the assembled contigs. Thirty-seven protein-coding genes were then annotated manually on the contigs, as well as 5S, 18S and 26S rRNA and tRNA genes. The vast majority of genes found in other plant species were found in the ash mitochondrial assembly (Fig. 4.7). A few genes that were absent in the ash mitochondrial assembly (e.g., *rps1, rps2, rps11, rps19*) are also absent in many other plant species, particularly in many

Table 4.1: Statistics of mitochondrial genome assembly, from the very first assembly to the most recent version after several rounds of map-extend-join.

	Initial assembly	**Current assembly**
Number of Contigs	296	26
N50	1.8 kbp	60 kbp
% gaps (N)	6.25	0.05
Longest contig	11 kbp	184 kbp

of the dicots. Therefore it is to be expected that these genes may not be annotated in ash.

I marked one gene as partially annotated in ash; *atp6* has an initial region of low conservation at the start of the gene (Fig. 4.8), making it difficult to identify the start location from synteny alone. The following region of high conservation aligned well to that of six other plant species, and the initial region of low conservation still aligned well to other Asterid species *Helianthus annuus* and *Mimulus guttatus*, but not quite as well to *Nicotiana tabacum*.

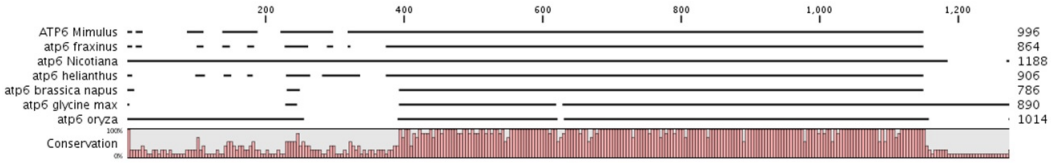


Figure 4.8: Alignment of *atp6* gene from six plant species shows an initial region of low conservation followed by a region of high conservation. The *Fraxinus* gene appears very similar to that of fellow Asterids *Helianthus annuus* and *Mimulus guttatus*, but less similar to *Nicotiana tabacum*.

In the same way as the plastid chromosome, I used the mitochondrial assembly in a BLAST search against the original *de novo* assembly, but the results were not quite as straightforward. Because most of the mitochondrial scaffolds are smaller than the plastid, and the assembly is not complete, mostly only small regions of the *de novo* assembly matched. In a few cases, a large part of the scaffold matched, and then the whole region was removed from the assembly (Contig7015, Contig6435, Contig2707, Contig2543, and Contig5453). The mitochondrial scaffolds were then added to the original assembly. Some regions may be duplicated where I could not identify their matches in the original assembly for certain. However, I prefer this scenario over the alternative of removing many small regions with uncertain matches, and inevitably breaking up many scaffolds in the process. In addition, there may be small regions of the nuclear genome that match the organellar genomes, and I did not want to delete these mistakenly.

Spurious matches between the organellar and nuclear assemblies could be caused by incomplete assemblies in either case, or alternatively they could be due to biological factors. Insertions and transfer of organellar DNA into the nuclear genome are well studied in plants [Ranade et al. 2016; Greiner & Bock 2013; Bock & Timmis 2008; Elo et al. 2003]. These occurred frequently in early plant evolutionary history due to double-strand break repair and non-homologous end joining [Kleine et al. 2009]. As a result, some organellar genes are

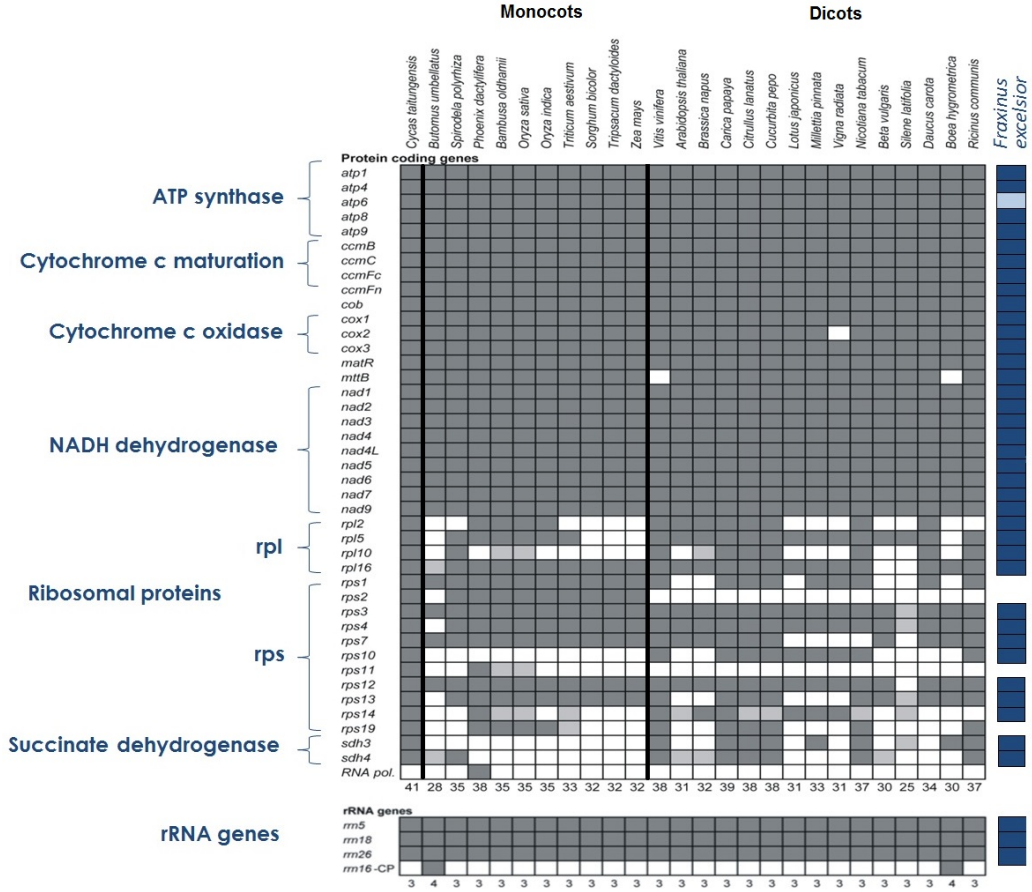


Figure 4.7: Genes annotated on mitochondrial chromosome assemblies of various plants, including ash (in blue). Dark colours represent complete annotation, lighter colour represent partial annotation, and white means complete absence of gene in mitochondrial genome. Only *atp6* is partially annotated in the ash assembly. Figure adapted from Cuenca et al. (2013).

now encoded in the nuclear genome. Any of the these processes make organellar assembly challenging, and therefore a different sequencing technology or organellar isolation may be needed to assemble these genomes completely.

4.3 Conclusion

Mitochondria and plastids are difficult to separate out from other cell contents and to sequence in isolation. I found a method of assembling and annotating the organellar genomes from whole genome sequencing reads to be successful. I was able to assemble the plastid genome into one circular chromosome and identify the inverted repeat section, and the mitochondrial genome into 26 scaffolds. The mitochondrial genome is more difficult to assemble than the plastid because it experiences recombination, can be bi-parentally inherited and may therefore contain heterozygosity [Davila et al. 2011; Barr et al. 2005]. Using alignments with other plant organellar genes, I was able to annotate 37 protein-coding genes as well as the 5S, 18S and 26S rRNA and tRNA genes on the mitochondrial scaffolds, and 72

protein-coding gene, 16S, 23S and 5S rRNA and tRNA genes on the plastid chromosome.

There are many uses of organellar genomes. They give different insights into inheritance and population structure than the nuclear genome, especially the plastid genome being only uni-parentally inherited (e.g., Heuertz et al. 2004b), as well as containing unique genes. They are also generally more conserved across groups of species than the nuclear genome, and can therefore be used for more broad phylogenetic studies (reviewed in Patwardhan et al. 2014), such as has been done using the chloroplast genes *ndhF*, *rpl16* and *matK* [Liang et al. 1996; Kim et al. 1995; Zhang et al. 2000]. I was also able to use a part of the plastid sequence in a range-wide study of population structure in European ash trees (see Chapter 6).

Further improvements, particularly in the mitochondrial genome, can be achieved by gaining additional sequencing data. There are still gaps in the mitochondrial scaffolds generated by the joining of contigs, and these could be filled using Sanger sequencing. Similar to the whole genome *de novo* assembly, the mitochondrial genome could also be improved by new long-read technologies such as PacBio and Oxford Nanopore (see Section 3.5).

Chapter 5

Analysis of whole genome duplications in *Fraxinus excelsior*

5.1 Introduction

5.1.1 A rich history of whole genome duplications (WGD) in plants

Genome duplications have occurred extensively throughout the plant kingdom, with estimations of polyploidy ranging from 30-70% of angiosperm clades [Soltis et al. 2015; Wendel et al. 2016]. In fact, evidence suggests that all eudicots share an ancestral WGD event, as do all angiosperms, and all seed plants [Jiao et al. 2011]. The phylogenetic tree of analysed plant genomes in Fig. 5.1 shows common WGD events (coloured squares) at the base of these lineages. Therefore, all extant seed plant species have, at one time or another, been polyploid, although many cannot be identified as such using chromosome counts when we study their genomes today. Only recent polyploids have detectable duplicated chromosome complements. Many of these, certainly the best-studied, are crop plants. For example, potato (*Solanum tuberosum*) and cotton are recent tetraploids (4x genome)[Li et al. 2014b; The Potato Genome Sequencing Consortium 2011], and bread wheat is a recent hexaploid (6x genome) [Brenchley et al. 2012]. Polyploidy has also been extensively studied in the model organism *Arabidopsis* [e.g., reviewed in Bomblies & Madlung 2014]. Whilst European ash does not have a polyploid chromosome complement, it was not previously known whether the ash lineage has had any WGD events (bar the core shared events). Therefore, the aim of this chapter is to search for evidence of WGD in the ash genome.

Some polyploids arise from the hybridisation of two closely related species. Such hybrids are usually, and should be by definition, sterile, as the homologous chromosomes cannot form bivalents during meiosis and therefore gametes do not form properly. However, non-reduction can occur during meiosis through meioetic defects such as abnormal spindle formation, disrupted cytokinesis, or the omission of one cell division altogether [Brownfield & Kohler 2011]. Non-reduced gametes have a diploid chromosome complement and can therefore form a polyploid zygote upon fertilisation; at this point the genome content is doubled [Madlung 2013; Levin 2013; Brownfield & Kohler 2011; Ramsay & Schemske 1998). Upon meiosis in the resulting polyploid plant, the polyploid genome behaves effectively as two diploid genomes; bivalents form and between the homologs within each parental genome,

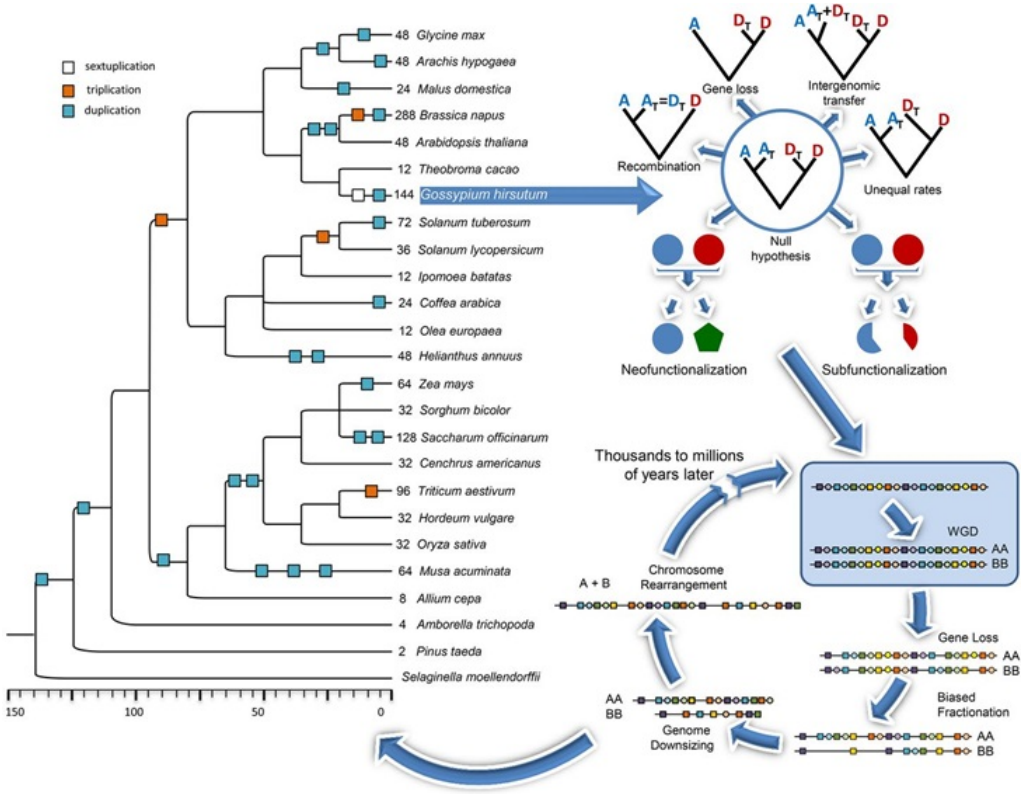


Figure 5.1: Patterns and processes of genome duplications in plants. Image copied from Wendel (2015). Left: A small fraction of WGD occurrences in plant kingdom. Numbers at branch tips indicate the number of genome equivalents derived by multiplication of all previous polyploidisation events. Top right: some genomic consequences to polyploidisation, with a particular focus on allopolyploids. DNA-level responses include intergenomic transfer (between the two homoeologous genomes, A and D), and gene loss, wheile expression-level responses include neofunctionalisation (functions of the two homeologs diverge completely) and subfunctionalisation (partial functions of the two homeologs are lost, so that only by being expressed together can they provide the function of the original gene). Bottom right: long-term responses after genome duplication, involving gene loss and fractionation towards a diploid genome again. The whole process is typically cyclical, with repeated polyploidisation events and subsequent diploidisation processes occurring several times in some lineages.

allowing proper disomic chromosome segregation. Polyploidy is commonly observed in hybrids, because in cases where divergence of the parents species occurred a long time ago (over 4-5 mya [Levin 2013]), the only process by which hybrid progeny will be fertile is via genome doubling. Well-studied examples of allopolyploids are the crop plants *Brassica napus* (oilseed rape) and *Triticum aestivum* (bread wheat). Autopolyploids on the other hand, are the result of a genome doubling within a single species which can occur spontaneously (as in potato) or again from the fusion of unreduced gametes. Instead of having two slightly different diploid genomes, autotetraploids possess four near identical copies of the haploid genome. In both cases, polyploidisation can result in irregularities during meiosis. Aberrant pairing in allopolyploids and multivalent complexes in autopolyploids can result in aneuploidy (an abnormal number of chromosomes, such as four chromosomes splitting into one and three during meiosis) and extensive chromosome rearrangements [Hollister 2015].

It has been well documented that many plant species have experienced not one WGD, but several repeated polyploidisation events. However, when considering the genome size and chromosome number of these species, it appears as if the content is not particularly in excess of ancestral angiosperm genomes (typically 5-7 chromosomes [Wendel 2015]), and certainly not reaching levels suggested by the branch numbers in Fig. 5.1, where these numbers indicate the number of theoretical genome equivalents derived by multiplication of all previous polyploidisation events.

Duplicated genomes undergo several processes that reduce gene and chromosome numbers down towards diploid levels. Immediately after a polyploidisation event, the genome enters a very dynamic state. The 'genomic shock' of doubling all content can result in several rapid dosage compensation mechanisms, which serve to reduce the expression of genes back down to diploid levels and re-stabilise the genome. These include large translocations, deletions, inversions, rRNA copy number variation, transcriptional modification, epigenetic remodelling (e.g., DNA methylation, histone modifications) and transposon activation [McClintock 1984; Song et al. 1995; Renny-Byfield & Wendel 2014]. Long-term processes that reduce gene and chromosome content occur at both the DNA sequence level and higher levels such as gene expression and protein function (Fig. 5.1). Sequence-level processes include mutation-induced gene silencing (e.g., introducing a premature stop codon) or gene loss altogether, transfer of genomic content through recombination (typically repetitive sequences such as transposable elements), and loss or further duplication of large segments through unequal recombination [Soltis et al. 2015; Wendel et al. 2015]. In addition, the resulting redundancy from WGD relaxes selection pressure on duplicate genes, allowing divergence of sequences and functions via accumulation of nucleotide substitutions, possibly at unequal rates. Higher level responses include expression dominance of one progenitor genome over the other [Leach et al. 2014; Buggs et al. 2010] (one potential mechanism is differential DNA methylation leading to unequal silencing), neofunctionalisation, where the two homeologs diverge so much as to develop different functions, and subfunctionalisation, where the two homeologs each lose one part of the protein's function (e.g., they each lose a different protein domain) so that the protein will only function correctly when both genes are expressed together. Over a long period of time, these processes can act unequally so that one of the progenitors is preferentially retained over the other; this 'biased fractionation' has been observed in many studies [Renny-Byfield et al. 2015; Freeling et al. 2012; Cheng et al. 2012; Paterson et al. 2012; Schnable et al. 2011a; Schnable et al. 2011b; Freeling 2009]. Genomes are eventually downsized and chromosomes rearranged sufficiently to appear again as if diploid.

Polyploidisation and subsequent diploidisation is a cyclical process. Fig. 5.1 shows that many plants have undergone several polyploidisation events in their history, where the polyploidisation-diploidisation cycle has been repeated many times. Evidence of ancient WGD may be difficult to detect after such long-term duplicate reduction, and older duplicates being masked by more recent WGD. However, more recent duplications can be detected easily by making use of the extant copies of homeologs and the divergence of their sequences using the Ks method.

5.1.2 Overview of the Ks method

The Ks value, or dN/dS, is the measure of synonymous substitutions per synonymous site between two sequences, where synonymous substitutions are those that do not change the protein sequence. This type of mutation is typically neutral, i.e., not under any selection pressure [Maere et al. 2005]. Ks is a useful measure of time since the divergence of two originally identical sequences since, without selection, mutations occur by chance at an average constant rate. Ks values can therefore be used as a molecular clock that indicates time since a duplication event occurred. As the mutation rate of each species differs, so does the rate at which paralogs diverge. Therefore Ks distributions cannot be compared definitively between different species unless they have similar mutation rates over time.

Ks values can be calculated directly from an alignment of two DNA coding sequences which also have codon information. Only with the codon information can polymorphisms in the alignment be classified as either synonymous or non-synonymous, and the synonymous substitution rate calculated. As there are usually many pairs or groups of homeologs in a genome, a great number of Ks values can be used to build a genome-wide distribution. A typical Ks profile without detectable WGD events resembles an 'L' shape, with many recent local duplications that have a small Ks, and fewer older duplications with a large Ks [Lynch & Conery 2000]. Duplicated genes are gradually lost over time, due to several processes mentioned previously that maintain gene dosage at normal levels, e.g., neofunctionalisation, subfunctionalisation, silencing and deletion. This loss of duplicates occurs, on average, at a steady rate of decay [Maere et al. 2005]. Events that duplicate the total genomic content cause a strong peak in the Ks distribution, where a greater number of paralogs have a similar Ks value than would be expected from the gradual decay of local duplications. Duplication events that occurred longer ago are less likely to still be represented by paralogs that are sufficiently similar to be aligned, and therefore the peaks at high Ks values tend to be lower and broader than recent Ks, as shown in Fig. 5.2.

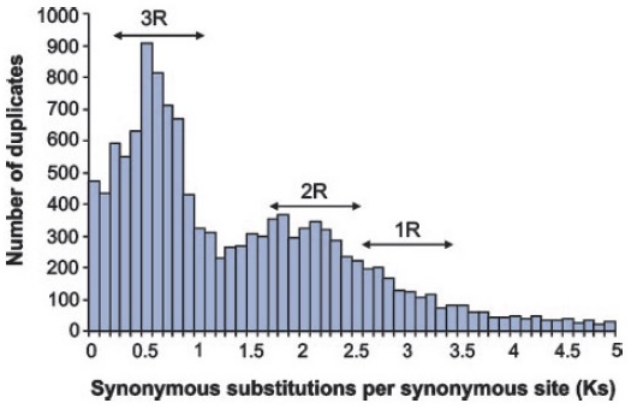


Figure 5.2: Ks distribution of *Arabidopsis thaliana*, with 3R, 2R and 1R representing three WGD events in its history. The most recent WGD (3R) is certainly the strongest peak, whereas the second peak (2R) is very low and broad, and the oldest WGD event (1R) is hardly even detectable. Older peaks tend to be lower and broader because of the gradual loss of homeologs. Image taken from Maere et al. (2005).

5.1.3 Correcting for redundant Ks values in homeolog groups

When Ks values are calculated for a group of paralogs, certain steps need to be taken in order to avoid over-inflating the number of Ks values that contribute to the genome-wide distribution. In brief, for a group of *n* paralogs, the number of single gene duplication events that gave rise to the group should be *n-1* [Maere et al. 2005; Blanc & Wolfe 2004]. However, by calculating Ks values for every pairwise combination of genes within a group, one obtains *n(n+1) / 2* measures. Many of these values are therefore redundant, and need to be corrected otherwise the total set of Ks values for one genome will be enriched for those from large paralog groups. One way to achieve this is described in Maere et al. (2005) and Blanc & Wolfe (2004). As a hypothetical example, consider a group of genes: A, B, C, and D. The smallest Ks value (representing the most recent duplication) is between genes A and B, which is then the first value to be taken as a true duplication Ks. Therefore genes A and B are grouped together. Ks values are then re-calculated for all other paralogs in the group with each other and with the collective group of the first two genes, i.e., C to D, and also C to (A and B), and D to (A and B). The Ks between C to (A and B) is simply the average of C to A and C to B (and the same goes for D). The smallest Ks value of this second round of calculation is the next to be considered a true duplication Ks. Two scenarios can occur at this stage; either one of C or D have the smallest Ks to (A and B), and these are therefore grouped together into a larger group. In this case, the last Ks to be calculated will be the remaining gene to the group of now three genes, e.g., D to (C + (A + B)). The second scenario, is that the Ks between C to D is smaller than either C to (A and B), and D to (A and B). In this case, the last Ks to be calculated will be the average of (A + B) to (C + D). Ultimately, three Ks values will have been calculated in this example, instead of six pairwise comparisons, and the three corrected Ks values will be added to the genome-wide distribution.

5.2 Methods

I examined the evidence for historical WGD event(s) in ash, and compared these with six other plant species: *Olea europaea* (olive), *Solanum lycopersicum* (tomato), *Mimulus guttatus* (monkey flower), *Coffea canephora* (coffee), *Utricularia gibba* (bladderwort) and *Vitis vinifera* (grape). The reasoning behind choosing these six additional species is as follows: olive is the most closely-related species to ash that has genome sequence information available; both species are members of the Oleaceae family. It is not yet known whether a WGD event is detectable in the olive genome. Monkey flower and bladderwort are members of the Lamiales order along with ash and olive. Bladderwort was found to have three detectable WGD events [Ibarra-Laclette et al. 2013], one of which is shared with monkey flower. If this shared event is common to both species, it could also be shared with other Lamiales members ash and olive. Tomato and coffee share only membership of the clade Asterids with ash, as tomato lies in the order Solanales and coffee in Gentianales. Tomato is known to have a genome triplication [The Tomato Genome Consortium 2012], whereas coffee has no detectable WGD event [Denoeud et al. 2014]. Finally, grape is separated from the other six species by residing in the Rosid clade of the angiosperms, and therefore being the least related to ash. However it was previously found that grape also has no detectable recent WGD

event in its history (except for the shared ancestral WGD events which are not detectable using the Ks method) [Jaillon et al. 2007; Denoeud et al. 2014]. The Ks plots of grape and coffee should not therefore show any peaks and will provide a baseline comparison for other species which may or may not show evidence of an historic WGD event. The phylogeny and WGD history of these species can be visualised in Figure 5.3 where ash and olive reside in the Lamiales branch along with *Mimulus* and *Utricularia*.

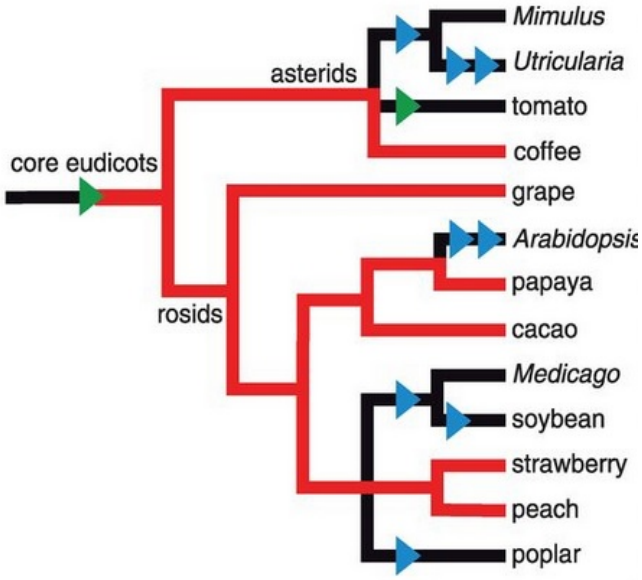


Figure 5.3: WGD event history in core angiosperms, with duplication events shown by blue triangles, and a triplication event shown by a green triangle. Image taken from Deneoud et al. (2014). Ash and olive reside in the Lamiales order along with *Mimulus* and *Utricularia*

I first obtained CDS and protein fasta sequences for all species. CDS and protein sequences were provided by David Swarbreck and Gemy Kaithokottil from TGAC, after their genome annotation on the BATG-0.5 assembly [Sollars et al. 2017]. CDS and protein sequence were downloaded from the Phytozome v10.3 repository (https://phytozome.jgi.doe.gov /pz/portal.html) for monkey flower, tomato, and grape. Coffee CDS and protein sequences were downloaded from the Coffee Genome Hub website, http://coffee-genome.org/download. Sequences for bladderwort were obtained by my QMUL colleague Laura Kelly from the author of Ibarra-Laclette et al. (2013). Olive open reading frames were predicted from trancriptome data [Munoz-Merida et al. 2013] by my QMUL colleague Endymion Cooper, using TransDeCoder v2.01 [http://transdecoder.github.io/].

All CDS models for each species were used in an all-against-all BLASTn search, with an E-value threshold of 1e-5. Results were then filtered using a custom Perl script to remove hits where each CDS matched with itself, and to remove any inter-gene hits with similarity less than 50% of the length of the longest gene. This ensures that genes match at least half of their length, and in doing so tries to avoid any matches that have just hit single protein domains. Next, groups of paralogous genes were collected using the following criteria: any one gene matching another one gene created a pair, any one gene matching one gene within a group of paralogous genes was added to the group, and any one gene already in a group matching another one gene in another group resulted in the two groups being

merged. Ks values were then calculated for each pairwise combination of genes within the group using a variety of software. Protein sequences were aligned using clustalw2 [Larkin et al. 2007] and were realigned with codon information using Pal2Nal v14 [Suyama et al. 2006]. This alignment was then run through clustalw2 again but only to output in Phylip format, which can then be read by the PAMLv4.8 program yn00. yn00 takes the alignment of two genes and calculates, amongst others, the Ks value. This was all performed within a custom Perl script. Ks values were corrected for redundant values among groups of paralogs using the python script RemoveRedundant.py, written by Endymion Cooper, which follows the method described in Maere et al. (2005) and Blanc & Wolfe (2004). All Perl and Python scripts (written by myself and Endymion Cooper, respectively) used in this analysis are available in my Github repository, https://github.com/lsollars/Perl-Scripts.

The complete set of corrected Ks values for all paralog groups and pairs in the genome was then plotted for each of the species using R. In addition, I also removed genes duplicated directly next to each other and re-plotted the Ks values. The GFF annotation file provided by TGAC was used to identify which genes were neighbours. These would likely represent local tandem duplications instead of being homeologs retained from WGD events, and should not therefore be considered when searching for peaks caused by WGD events in the Ks plots.

5.3 Evidence for two WGD events in the ash lineage

Ks plots for the seven species are shown in Figure 5.4 (end of chapter), as well as Ks plots with tandem duplications removed, which are shown as the right-hand graphs in each case.

The first important finding from this analysis is that the ash Ks plot shows evidence of two historical WGD events which look to be shared with olive. Both species have a large peak at Ks 0.2-0.3, and a small shoulder at Ks 0.6-0.7. It is to be expected that the older WGD shows a lower, broader peak than the more recent event [Maere et al. 2005]. We cannot be completely certain that the WGD event is shared, as divergence between paralogs (and thus the KS values) depend on the mutation rate, which could be different in ash and olive. As there is no prior research on WGD in olive, I cannot compare my results with that of previous findings. However, as both species share the same peaks, the findings support each other and it would therefore make sense to conclude that even the most recent WGD event occurred before the ash and olive lineages diverged. I would predict that other species in the Oleaceae family would also show the same peaks. In addition, due to the lack of genome annotation for olive, I could not determine which paralogs were local duplications and therefore could not make any efforts to remove the 'noise' from the plot. I would expect that the older WGD peak in olive would become more pronounced if tandems were removed, which can be seen in some other species in Fig. 5.4, particularly tomato and monkey flower.

The first peak at 0.2-0.3 is much higher and sharper in ash than in olive, and indeed the numbers of genes identified as paralogs overall is much higher in ash than in the other six species. This might suggest that ash has many more genes retained from its most recent duplication event. However, when considering the number of transcripts in each species to

begin with, it can be seen that the number in ash ash (43,298) is much higher than that of the other six species (34,727 in tomato, 28,140 in monkey flower, 26,346 in grape, 25,574 in coffee and 28,494 in bladderwort). This could be an artifact from under-assembled heterozygosity during the assembly process, in which similar haplotypes are retained in different contigs. If genes are annotated in these contigs they would appear as paralogs with some nucleotide substitutions when aligned against each other. The gene annotation process could have also identified erroneous transcripts that were not filtered out in later QC steps, increasing the starting number of genes to analyze. In addition, the relatively low e-value threshold used in the BLAST search could allow more noise in the data and elevate the numbers of genes classed as paralogs, however this threshold was the same for each species analyzed and would not increase the number of ash genes alone. Numbers of genes belonging to each paralog group reported in Sollars et al. (2017) (the analysis performed by Endymion Cooper using a protein sequence BLAST) show similar elevated levels in ash compared to the other six species, indicating that the starting data (the gene annotations) are causing the elevation rather than the analysis method.

Tomato has a strong peak at Ks 0.6 which becomes much more defined when tandemly-duplicated genes are removed. Tomato was found to have a triplication event in its recent history [The Tomato Genome Consortium 2012]. Therefore, this event is not thought to be shared with that of ash and olive. The Ks plot from the genome paper is shown in Fig. 5.5; which has an almost identical peak at Ks 0.6. The Ks values in the tomato genome paper were calculated from syntenic regions in the genome, rather than from all possible paralogs, which could explain the low density of paralogs at very low Ks values in this figure. By considering only large syntenic regions, the authors ensure that the paralogs they identify originate largely from true WGD events (therefore considered homeologs) and not local duplications of single genes.

The Ks plot for monkey flower does not quite show a detectable peak in the first Ks plot in 5.4, however when tandemly-duplicated genes are removed, there is a small peak centering around 0.8-0.9. The Ks plot from the tomato genome paper (Fig. 5.5) also shows that monkey flower (blue line) has a gently sloping peak centered around 0.8-0.9. It is also shown to have a shared WGD event with bladderwort in the phylogeny figure (Fig. 5.3).

The Ks plots for bladderwort show quite different results to those expected from Ibarra-Laclette et al. (2013). In the first Ks plots in Fig. 5.4, bladderwort shows no definite sign of WGD peaks until a very high ks value of around 2.7, however when tandems are removed there is a slight shoulder around 0.4. The genome paper identifies two WGD events in addition to an older event thought to be shared with monkey flower, meaning that the *Utricularia* genome content is currently 8x that of the paleohexaploid core eudicot ancestor. I did not identify these events using my method. Bladderwort was found to have high rates of gene loss compared to tomato where many syntenic genes are only present now in a single copy, as well as a low amount of repetitive DNA [Ibarra-Laclette et al. 2013]. An alternative Ks plot is presented in Sollars et al. (2017) which shows a bladderwort Ks plot which agrees with Ibarra-Laclette et al. 2013. This Ks plot was generated using a protein BLAST instead of CDS.

Fig 5.3, taken from the coffee genome paper [Deneoud et al. 2014], shows that the tomato triplication is separate from the duplication event common to monkey flower and bladderwort, as well as the two WGD events specific to bladderwort. Although the Ks values for the peaks in monkey flower, ash and olive cannot be directly compared due to likelihood of having different mutation rates, it is possible that the WGD event shared between monkey flower and bladderwort is also shared with olive and ash. The peak in monkey flower is at a slightly higher Ks value than in ash and olive (0.8-0.9 compared to 0.6-0.7); this could be due to a shorter generation time in the herbaceous plant leading to a higher mutation rate. This WGD event could therefore be common to all the Lamiales order.

Neither grape or coffee showed signs of historical WGD events, only a logarithmic decay in the frequency of Ks measures at high values. The higher frequency of paralogs at very low Ks values likely arise from recent local duplications that quickly diverge in their sequence. There is no single peak that indicates a lot of duplications occurred at the same time, like there are in the previous species. It was already known that recent WGD events are not detectable in the grape genome [Jaillon et al. 2007]. The genome paper of coffee also found no evidence of the genome triplication event found in tomato or the duplication event common to monkey flower and bladderwort [Deneoud et al. 2014]. An updated phylogeny of these seven species, with the new WGD event shared by ash and olive is shown in Fig. 5.6.

5.4 Conclusion

In conclusion, I have identified two recent WGD events in ash, the most recent of which seems to be shared with olive and could be common to all Oleaceae. This finding provides new information to the knowledgebase of WGD in plants, adding another WGD event to the growing list of known genome duplications. The older WGD event could be shared with monkey flower and bladderwort, possibly common to all species in the Lamiales order. However there are uncertainties regarding the timing of the WGD event due to a lack of knowledge of each species' mutation rates. The genome triplication in tomato however, is a separate event and not shared with any species studied here. The final member of the Asterid clade studied, coffee, did not show signs of a WGD event (therefore none of the detected events are common to all Asterids), nor did Rosid species grape. My findings are largely complementary to those of previous studies [Deneoud et al. 2014; The Tomato Genome Consortium 2012].

The wider implications of having two recent WGD events are that much of the ash genome content will still be duplicated. Although many processes act to reduce the gene content back to diploid levels (see section 5.1.1), there are clearly many genes remaining with at least two copies as these cause the peaks in the Ks plots. Future genome analyses should take this duplicated content into account. For example, mapping reads could result in a large amount mapping equally well to different locations in the genome, which could consequently affect variant calling. In some variant callers (depending on user-defined settings) non-unique reads are ignored. Alternatively, variants can be called but it can be difficult to

confirm to which copy of the gene they truly belong. This is also a well-known consideration when genotyping pseudogenes. When variants or expression values are used in association studies to identify potential markers, erroneous results can ensue. One solution is to use long read technology or mate paired reads with a long insert size. This would increase the chance of half of the read / one member of the pair mapping to a unique (non-duplicate) region, and therefore the read or pair can be mapped uniquely.

As ash has a diploid chromosome complement, it has obviously lost much of its duplicated content through processes discussed in section 5.1.1 such as deletion and chromosome rearrangements. Many diploidisation processes also act on genes, such as sub - and neofuctionalisation, as well as silencing through epigenetic pathways. Silencing can often occur unequally, so that one copy of a duplicated gene is kept active while the other is silenced. This will be investigated in the ash genome using DNA methylation data in Chapter 7.

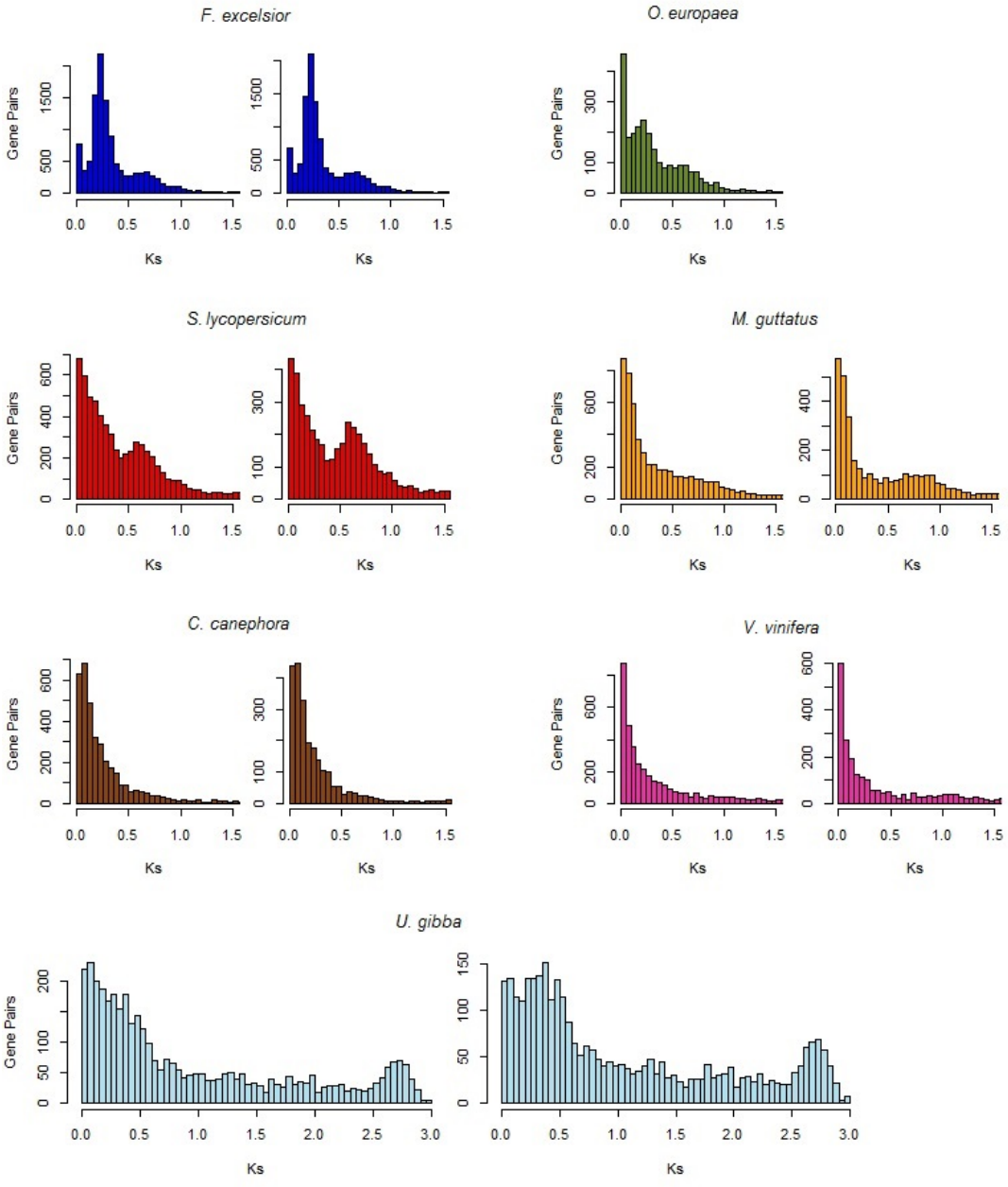


Figure 5.4: Ks plots for seven plant species using all paralogous genes (left-hand images), and with tandemly-duplicated genes removed (right-hand images). As a structural annotation for *Olea europaea* was not available, the Ks plot with tandems removed could not be drawn for this species.

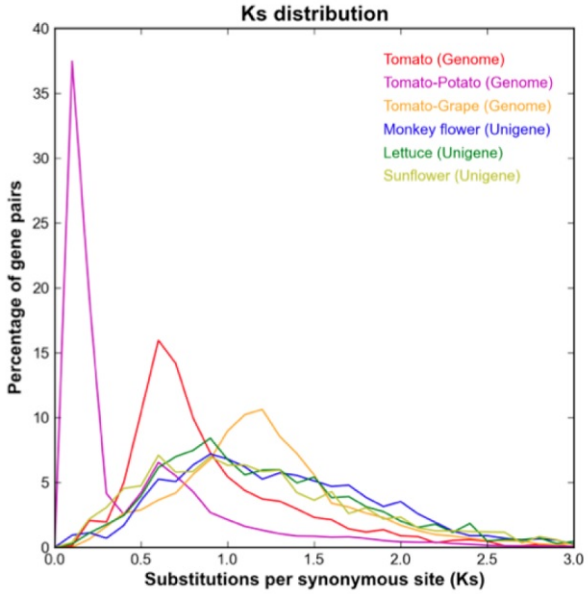


Figure 5.5: Distributions of synonymous nucleotide substitution (Ks) rates between syntenic regions for tomato, potato, monkey flower, lettuce and sunflower genomes. Image taken from The Tomato Genome Consortium (2012) Supplementary Information.

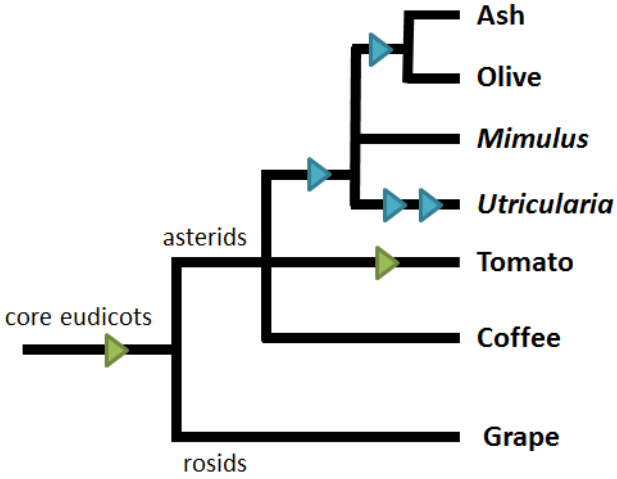


Figure 5.6: Updated WGD phylogeny of the seven species studied, showing the newly discovered WGD event shared between olive and ash. Blue triangles show genome duplications and green show triplications

Chapter 6

Population structure among European ash trees

6.1 Introduction

6.1.1 Current population research on ash

Current genetic population structure is shaped by historical colonisation trends, climate variation, population fragmentation, adaptation to environmental variables and movement of individuals by humans. A common finding in the phylogeography of European tree species are diversity hotspots due to refugia in Southern Europe during the Last Glacial Maximum (LGM), primarily in the Iberian, Italian and Balkan peninsulas [Petit et al. 2003], with some additional refugia identified in the Carpathian and Apennine Mountains. These areas have been identified as refugia of numerous forest tree species including *Populus* spp. [Fussi et al. 2010], sweet chestnut (*Castanea sativa*) [Mattioni et al. 2013; Poljak et al. 2017], oaks (*Quercus* spp.) [Petit et al. 2002; Brewer et al. 2002], Scots pine (*Pinus sylvestris*) [Wjkiewicza & Wachowiak 2016] and silver fir (*Abies alba*) [Liepelt et al. 2009; Piotti et al. 2017].

Several previous studies have examined population structure in *F. excelsior* using genetic markers. However, results differ among studies and the type of marker used. I will first review results using nuclear markers and then compare them to those found using chloroplast markers. The largest differences between these types of markers are that firstly, nuclear markers are inherited bi-parentally and chloroplast markers only through the maternal line. Secondly, much of the chloroplast sequence is very conserved and therefore nuclear markers are much quicker to evolve and diversify, especially at neutral loci which are not under selection pressure.

Only a few studies have assessed population structure across most of the European range of *F. excelsior*. Studies using nuclear microsatellites tend to find one very large diverse population in western and central Europe [Heuertz et al. 2004a; Tollesfrud et al. 2016]. Because *F. excelsior* is wind-pollinated and has an outcrossing mating system, extensive gene flow over large distances can cause this diverse panmictic population. An increase in population structure has been found in south-eastern Europe where many separate separate demes were

identified [Heuertz et al. 2004a; Tollesfrud et al. 2016]. These results support the hypothesis of glacial refugia in the Balkans where the majority of diversity still exists. Magyari et al. (2013) found fossil pollen evidence to suggest *F. excelsior* existed in the Carpathian mountains during the Last Glacial Maximum (LGM). Subsequently, post-glacial colonisation of central and western Europe occurred with high gene flow, resulting in homogeneous populations across these areas. Similar results were found in Irish [Beatty et al. 2015] and British ash [Sutherland et al. 2010]; these studies also found extremely low levels of population genetic differentiation using nuclear microsatellites, suggesting a single diverse meta-population across the UK, western and central Europe. However, Tollesfrud et al. (2016) found that ash in Northern Europe was distinct from that in central Europe, and in fact similar to South Eastern populations. The northern ash could have therefore arisen through a separate colonisation route from the Balkan peninsula, through Eastern Europe and to the north.

A different pattern is found using chloroplast RFLP (Restriction Fragment Length Polymorphism) and microsatellite markers [Heuertz et al. 2004b; Sutherland et al. 2010; Tollesfrud et al. 2016]. Heuertz et al. (2004b) found a much higher level of population differentiation and haplotype frequencies that show a strong geographical distibution (Fig. 6.1). The results of Tollesfrud et al. (2016) are very much in accordance with this. Using these patterns, they suggested post-glacial colonisation routes for ash from refugia in Italy to France and southern Germany, from Iberia to the UK, from the eastern Alps to Germany, Denmark, the Czech Republic and Poland, and from the Balkan Peninsula north into Ukraine, Russia, the Baltic states and also Sweden. Tollesfrud et al. (2016) also suggest the colonisation route from the Balkan peninsula through Eastern Europe and up to Scandinavia (excluding Denmark). They find evidence for a contact zone between two different colonising lineages in Northern Poland, where two different chloroplast haplotypes are now found. This can also be observed in Fig. 6.1, where the green and red haplotypes meet. Sutherland et al. (2010) also found evidence for a colonisation route for ash from Iberia to the UK, and found much more population structure using chloroplast microsatellites than nuclear markers. Heuertz et al. (2004b) explain the difference in results commonly seen between nuclear versus chloroplast markers as being due to poor mixing among post-glacial recolonising lineages, but extensive pollen flow between the different colonising populations. As chloroplast DNA is only inherited maternally through seeds which are not dispersed as far as pollen, high pollen flow will transmit nuclear DNA but not chloroplast DNA, producing homogenous nuclear DNA but maintaining localised chloroplast haplotypes.

Hybridisation between *F. excelsior* and fellow European species *F. angustifolia* has also been a common topic of study, as the ranges of the two species overlap throughout much of France, as well as northern Spain and central Italy, forming an easily studied hybrid zone where hybrid trees frequently occur [Gerard et al. 2013]. Patches of *F. angustifolia* can also be found as far north as Germany and east to the Balkans (Fig. 6.2). Hybrids can be easily distinguished using microsatellite markers, in combination with morphological traits [Thomasset et al. 2011]. Gerard et al. (2013) suggested that hybridisation occurred *after* the post-glacial expansion of *F. excelsior* and *F. angustifolia* ranges, due to current stable hybrid populations being maintained by winters with little frost and high summer

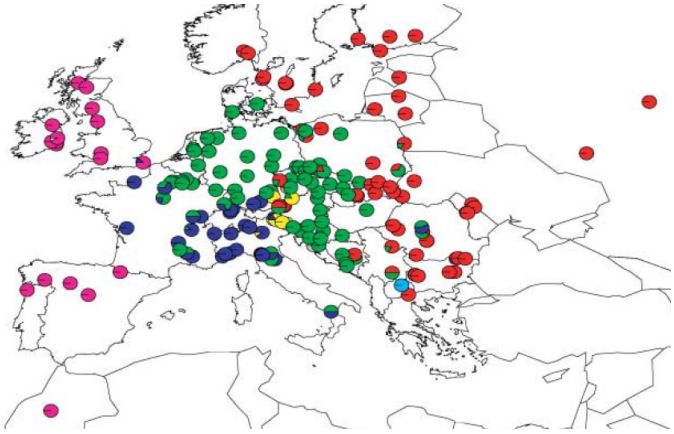


Figure 6.1: Geographical distribution and frequency of twelve haplotypes (colours on pie charts) identified using chloroplast microsatellite markers. Image from Heuertz et al. (2004b)

precipitation. Heuertz et al. (2006) also found high levels of haplotype-sharing between the two species but suggest hybridisation occurred in refugia *during* glaciation and also during post-glacial colonisation of central and northern Europe. Recent reforestation programmes have increased the movement of ash trees, particularly from continental Europe to the UK to re-stock plantations. This has resulted in gene flow from Europe, and in some cases, the introduction of hybrids, e.g. in Ireland [Thomasset et al. 2013]. Therefore, when trying to explain population structure patterns, potential recent movement by humans could be a contributing factor and should be taken into account. No haplotype-sharing was found for another European ash, *F. ornus*, with the other two species.

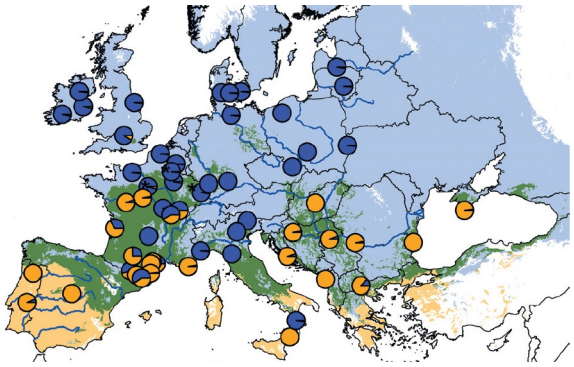


Figure 6.2: European distribution of *F. excelsior* (blue areas) and *F. angustifolia* (orange areas) and their range overlap (green areas). Pie charts show proportions of gene pools belonging to each species. Image taken from Gerard et al. (2013)

One drawback of all the aforementioned studies is that the number of markers used is very low, typically a few SSRs or microsatellites, or chloroplast DNA alone which is only inherited maternally. Organellar DNA can only provide insights into a single genealogy, while nuclear markers are inherited bi-parentally, are subject to recombination events, and therefore incorporate more historical processes in their record. Using a small number of markers also falls short of providing the total picture of evolutionary events. Different regions in the genome could have arisen from different historical population processes and therefore an increase in the number of markers provides a much more robust way of measuring population structure.

It is well-known that the cost of sequencing has fallen recently and will likely continue to do so. This has allowed the sequencing of genome-wide markers in increasingly large samples, of which one method is whole genome sequencing. Sequencing targeted regions has also increased in popularity within ecological and evolutionary research, using techniques

such as Restriction site Associated DNA-seq (RAD-seq). This involves sequencing of thousands of loci at deep coverage, and allows the genotyping of hundreds of individual samples. Therefore there will likely be more studies on ash population structure in the near future, using tens or hundreds of thousands of genome-wide markers in a large number of trees across the European range.

This chapter will describe investigations into population structure of European ash trees, using a range-wide diversity panel of 38 trees. We performed whole genome sequencing of 37 trees at shallow coverage (approximately 10x), to obtain nearly 400,000 genome-wide SNP markers. Three methods are used to interpret population structure; Principal Components Analysis (PCA), the software program STRUCTURE v2.3.4 [Pritchard et al. 2000], and haplotype networks of plastids. In the latter part of the chapter, I examine the effective population size history using this group of samples with two methods; PSMC [Li & Durbin 2011] and SNeP [Barbato et al. 2015].

6.1.2 Approaches for analyzing population structure

In this study I use three different methods to investigate population structure in European ash; the STRUCTURE software [Pritchard et al. 2000], Principal Components Analysis (PCA), and plastid haplotype networks. PCA is a common statistical tool for multidimensional data that tries to reduce the variability of the measures down into just a few main dimensions. STRUCTURE is an open-source program developed to cluster individuals into populations based on genomic marker data. The input data for both PCA and STRUCTURE are the same in this study; a set of nearly 400,000 polymorphic loci and their genotype information for each sampled ash tree (see Methods). However, the assumptions for these two methods are different and therefore the population structure results may differ also. PCA makes no assumptions about the data; it simply tries to represent the data in a way which maximizes the amount of variance explained. In contrast, the STRUCTURE program has many assumptions for its model, such as an expected amount of admixture, and that individuals belong to populations in Hardy-Weinberg equilibrium. The data for the plastid haplotype network is very different; a sequence alignment of the long single copy region of the plastid chromosome (assembled in Chapter 4) that clusters individuals based on the number of nucleotide substitutions between sequences. Therefore, I have chosen a mixture of distance-based and model-based methods, which could reveal different patterns in the population structure of the samples.

6.1.3 Approaches for estimating past N_e

There are three main methods for estimating N_e history from genomic data; Site Frequency Spectrum (SFS) or Allele Frequency Spectrum (AFS) methods, haplotype methods such as Time to Most Recent Common Ancestor (TMRCA), and Linkage Disequilibrium (LD) methods. Without the use of genomic data, inference of N_e is restricted to very few domestic populations where demographic measures have been recorded for a large number of generations [Barbato et al. 2015].

SFS uses the frequency of alleles found at neutral polymorphic loci to estimate changes

in N_e within one population or between multiple populations, by counting how many times alleles are shared among individuals or populations. The observed SFS is compared to the expected SFS from various demographic models of population history, and likelihood methods are used to choose the model that best fits the data. A couple of commonly used methods include *dadi* [Gutenkunst et al. 2009] and *fastsimcoal2* [Excoffier et al. 2013]. However, these methods have some drawbacks; they are often time-consuming and computationally intensive due to the need to perform many simulations, and may require many re-runs if several different model parameters need to be tweaked. SFS obtained from low coverage data can also miss rare alleles as these need to be present in at least a few individuals to be considered reliable and rule out sequencing errors [Excoffier et al. 2013]. These missing data can lead to inaccuracies in population inference. In addition, Myers et al. (2008) show that population history cannot always be inferred correctly from allele frequency data. Therefore, I did not implement any of these methods.

TMRCA approaches measure the TMRCA between the two alleles at each locus across the whole genome, and use the distribution of times to estimate the history of N_e. Each locus will have its own TMRCA between the two alleles, therefore by using a genome sequence with millions of loci, a large number of comparisons can be obtained. The program PSMC [Li & Durbin 2011] implements this approach by measuring the change in local density of heterozygous sites across the genome, on the basis that loci with two different alleles represent a TMRCA further in the past than a homozygous site. Loci located close to each other should also have fairly similar TMRCA measures, reflecting runs of heterozygosity or homozygosity. The genome is therefore partitioned into segments of local TMRCA similarity using HMM models, where a change in segment would in theory be caused by a historical recombination event (Fig. 6.3). The density of recombination events also reflects the effective population size, as low recombination suggests a small N_e, and vice versa. The distribution of heterozygote density is used to construct ancient N_e.

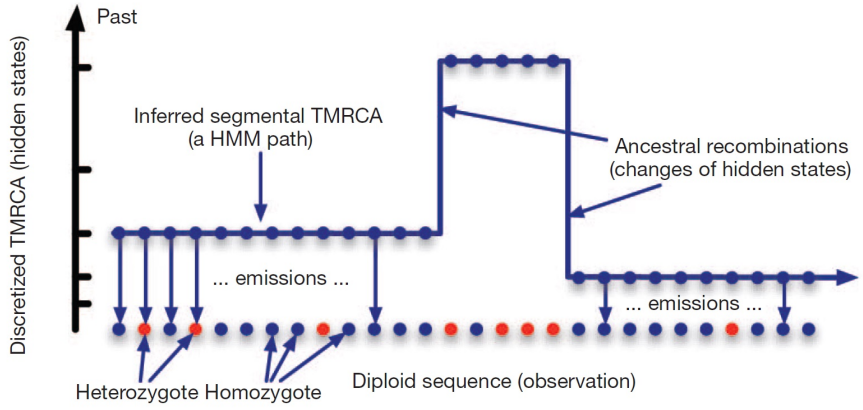


Figure 6.3: PSMC method: A Hidden Markov Model (HMM) is used to segment the genome into discrete TMRCA categories. TMRCA is inferred from the density of heterozygotes within each segment. Image taken from Li & Durbin (2011)

The final method uses LD between pairs of SNPs across the genome to infer historical N_e. Two loci can be in LD if they reside closely together, reflecting a lack of recombination

between the two sites. However, factors such as admixture, selection and genetic drift can also cause LD between sites located far away from each other in the genome. The value of LD between proximally located sites lends information on recent N_e, while the further away two SNPs are, the more informative they are about older N_e. SNeP [Barbato et al. 2015] is a program that uses LD to estimate historical N_e, by calculating the LD between SNPs in multiple individuals over varying distances in the genome. By putting these distances into bins, SNeP produces average N_e estimations for discrete time points in the past using pairwise comparisons of thousands of SNPs.

6.2 Methods

6.2.1 Locations and origins of samples

In the early 2000's, thirty seven *F. excelsior* trees were selected from natural populations across Europe to be part of a breeding trial in an EU-funded project named Realising Ash's Potential (RAP). The trees were planted at The Earth Trust's Paradise Wood in Oxfordshire, UK, in 2004. These 37 trees were sampled for sequencing in 2014. In addition, we obtained DNA-seq reads from "Tree35", a Danish tree found to have very low susceptibility to ADB during exposure and inoculation trials [Mckinney et al. 2011; McKinney et al. 2012]. Tree35 was sequenced by researchers at The Genome Analysis Centre (TGAC), UK, as part of the Open Ash Dieback (OADB) project. This addition makes a total of 38 trees for this diversity panel, with Tree35 being given the sample number '38'. The original locations of the trees are shown in Fig. 6.4 and Table 6.1. Unfortunately, the records for sample number 34 were lost and/or confused with others, and therefore its source location is uncertain.

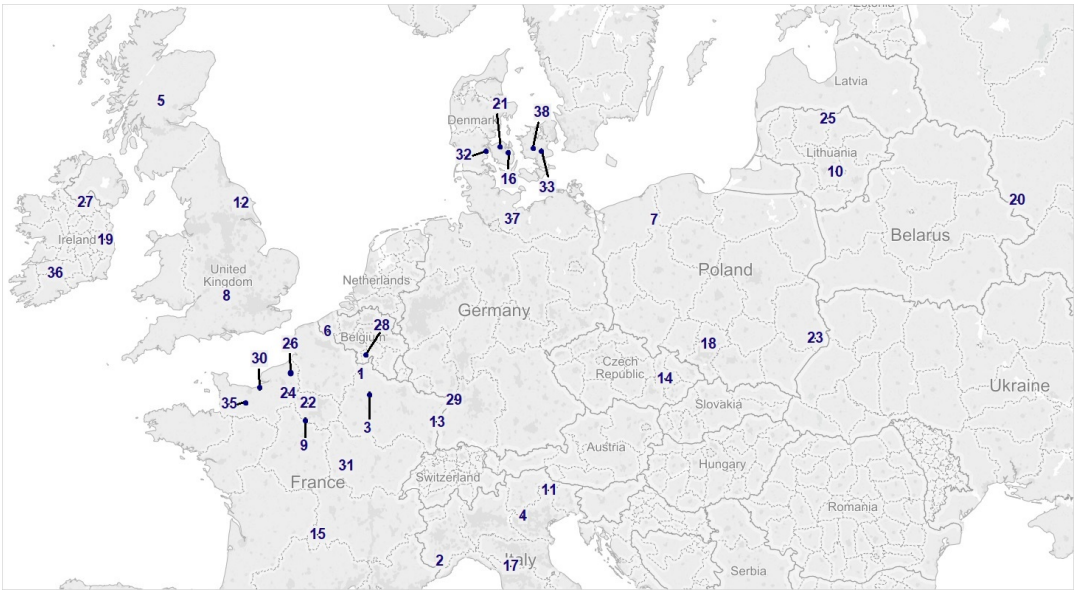


Figure 6.4: Original locations of 38 ash trees used in further population analyses. Sample number 34 is not shown as the original source location is uncertain

Table 6.1: Source locations of 38 trees used in population analyses, as well as sequencing, quality filtering, and mapping results. Average genome coverage shown in brackets (of 880 Mbp for 'raw' and 'filtered' reads, and of 714.8 Mbp non-N genome for 'mapped'

Sample Number	Source Location	Lat/Long	Millions of reads		
			Raw	Filtered	Mapped
1	La Romagne, France	49.69/4.29	44.75 (7.68x)	40.31 (6.18x)	32.26 (6.10x)
2	Valle Pesio, Italy	44.33/7.67	78.65 (13.50x)	68.05 (10.30x)	54.86 (10.24x)
3	Athis, France	49.02/4.61	54.77 (9.40x)	50.05 (7.67x)	40.87 (7.72x)
4	Monti Lessini, Italy	45.67/11.17	58.50 (10.04x)	53.89 (8.25x)	44.14 (8.33x)
5	Loch Tay, UK	56.57/-4.06	62.17 (10.67x)	57.06 (8.75x)	46.35 (8.76x)
6	Hoge Bos, Belgium	50.83/2.95	57.66 (9.89x)	49.57 (7.58x)	40.01 (7.55x)
7	Szczecinek, Poland	53.70/16.68	59.69 (10.24x)	52.13 (8.04x)	42.60 (8.10x)
8	Wytham Woods, UK	51.77/-1.33	52.28 (8.97x)	46.74 (7.20x)	38.52 (7.31x)
9	Dourdan, France	48.51/1.97	47.92 (8.22x)	38.80 (5.98x)	31.87 (6.06x)
10	Kaisiadorys, Lithuania	54.89/24.37	70.61 (12.12x)	60.99 (9.43x)	49.28 (9.40x)
11	Cadore, Italy	46.42/12.25	53.78 (9.23x)	50.38 (7.80x)	41.42 (7.91x)
12	Settrington, UK	54.12/-0.71	71.41 (12.25x)	66.78 (10.34x)	54.50 (10.42x)
13	Huttenheim, France	48.37/7.53	81.45 (13.98x)	71.94 (10.70x)	59.52 (10.91x)
14	Rabstejn, Czech Rep.	49.56/17.15	60.21 (10.33x)	50.97 (7.56x)	41.97 (7.67x)
15	Saint-Paul-de-Salers, France	45.12/2.53	61.41 (10.54x)	53.86 (8.00x)	44.10 (8.07x)
16	Ravnholt, Denmark	55.26/10.58	72.69 (12.47x)	63.48 (9.41x)	51.93 (9.49x)
17	Abetone, Italy	44.17/10.67	71.21 (12.22x)	62.24 (9.26x)	50.60 (9.28x)
18	Wloszczowa, Poland	50.51/19.01	86.04 (14.76x)	74.01 (10.98x)	60.76 (11.11x)
19	Donadea, Rep. of Ireland	53.21/-6.45	60.52 (10.38x)	51.45 (7.66x)	42.44 (7.80x)
20	Smolensk, Russia	54.20/32.00	89.96 (15.44x)	79.74 (11.95x)	65.49 (12.10x)
21	Stjernebjerg, Langesø, Denmark	55.44/10.19	77.90 (13.37x)	61.26 (9.15x)	50.12 (9.22x)
22	F. D. de Marly, France	48.89/2.04	65.46 (11.23x)	55.16 (8.28x)	45.30 (8.39x)
23	Mircze, Poland	50.65/23.50	73.19 (12.56x)	63.78 (9.55x)	52.47 (9.69x)
24	Le Hazey, France	49.16/1.28	59.65 (10.24x)	52.57 (7.87x)	43.49 (8.03x)
25	Zeimelis, Lithuania	56.16/24.02	48.47 (8.32x)	42.04 (6.40x)	34.28 (6.44x)
26	Monterolier, France	49.63/1.36	76.69 (13.16x)	67.32 (10.26x)	55.34 (10.40x)
27	Enniskillen, UK	54.14/-7.28	62.66 (10.75x)	56.84 (8.67x)	46.63 (8.78x)
28	Bois de Rose, Belgium	50.23/4.72	72.88 (12.51x)	65.42 (9.95x)	53.74 (10.08x)
29	Karlsruhe, Germany	48.98/8.30	65.36 (11.22x)	56.35 (8.56x)	46.00 (8.62x)
30	Saint-Gatien, France	49.35/0.14	66.73 (11.45x)	58.89 (8.99x)	47.99 (9.04x)
31	Aunais-en-Bazois, France	47.12/3.71	51.59 (8.85x)	36.96 (5.57x)	29.79 (5.56x)
32	Haderslev Østerskov, Denmark	55.29/9.54	49.90 (8.56x)	43.01 (6.53x)	34.78 (6.52x)
33	Bregentved, Denmark	55.34/11.96	59.79 (10.26x)	48.14 (7.29x)	39.13 (7.30x)
34	Uncertain	n/a	64.44 (11.06x)	52.99 (8.04x)	42.82 (8.02x)
35	Vassy, France	48.86/-0.70	54.51 (9.35x)	49.41 (7.54x)	40.07 (7.54x)
36	Curraghchase, Republic of Ireland	52.36/-8.53	59.89 (10.28x)	53.78 (8.20x)	43.93 (8.25x)
37	Herzogtum Lauenburg, Germany	53.73/10.72	48.05 (8.24x)	42.49 (6.47x)	34.62 (6.50x)
38	Tree35, Sorø, Denmark	55.39/11.59	43.75 (12.48x)	38.53 (9.77x)	30.07 (9.40x)

6.2.2 DNA sequencing and variant calling methods

Twig material was collected from the 37 RAP trial trees in 2014, and DNA was extracted from cambial tissue by Laura Kelly at QMUL, using a CTAB protocol [Doyle & Doyle 1987]. DNA was sequenced on an Illumina HiSeq at TGAC. DNA reads from Tree35 were downloaded from the OADB ftp site, now hosted at https://geefu.oadb.tsl.ac.uk/. Raw reads were trimmed using the CLC Genomics Workbench v7.5 to a minimum quality score of 0.01 (equivalent to Phred quality score of 20), a minimum length of 50 bp, and were also trimmed of any adaptor and repetitive telomere sequences. Filtered reads were mapped to the reference assembly using the 'Map Reads to Reference' tool in the CLC Genomics Workbench v8.0, setting both similarity match and length match parameters to 0.95. Filtering and mapping results are shown in Table 6.1.

My aim was to achieve a set of loci where all samples were sufficiently covered, to be able to have a complete set of genotype information without missing data. Therefore, regions

with coverage of between 5 and 30 reads in all samples were extracted using the 'Create Mapping Graph', 'Identify Graph Threshold Areas' and 'Calculus Track' tools in CLC Genomics Workbench v8.0. These extracted regions totaled 20.6 Mbp (2.3% of the genome).

Variant calling was performed on a read mapping pooled from all samples, using the 'Low Frequency Variant Caller' tool in the CLC Genomics Workbench v8.0, with the coverage-restricted regions from the previous step used as target regions. This prevented variants being called where some samples did not have read coverage, and also in the organellar scaffolds where the read coverage is very high. The following parameters were changed from default: Ignore positions with coverage above = 1000, Ignore broken pairs = no, Ignore non-specific matches = Reads, Minimum Coverage = 190 (38 samples with at least 5 reads each should have a combined total coverage of >190), Minimum Count = 10, Minimum Frequency = 5%, Base Quality Filter = Yes, Neighbourhood radius = 5, Minimum Central Quality = 20, Minimum neighbourhood quality = 15, Read Direction Filter = yes, Direction Frequency = 5%. As a result 529,812 variants were called, comprising 468,237 SNPs, 14,850 equal replacements (where >1 nucleotides are replaced by an equal number of nucleotides), 26,043 deletions, 19,085 insertions, and 1,597 unequal replacements (where at least one SNP lies directly beside an indel). The average quality of all reads at these variant positions was 36.2.

To genotype each sample individually at the variant loci called in the previous steps, the 'Identify Known Mutations from sample mappings' tool within the CLC Biomedical Genomics workbench v2.1 was used. The workflow takes a track of known variants as input (such as those called from the pooled read mapping) and reports the presence, absence, coverage, count and other statistics, of each variant locus in the read mapping of another sample (in this case, the read mapping from each of the 38 trees). The 'Identify Candidate Variants' tool was then used to filter variants with a minimum coverage of 5, minimum count of 3 and minimum frequency of 20%. VCF files for each tree were exported from the CLC Workbench and merged into one file using the vcf-merge tool from vcftools v0.1.12. The merged VCF file was then filtered using vcftools v0.1.12, to remove indels, multi-allelic loci, and loci with a Minimum Allele Frequency (MAF) <0.05, with 394,885 SNP loci remaining. This set of high quality SNPs with comprehensive knowledge of the genotype of every sample is used as input into population software STRUCTURE and for Principal Components Analysis (PCA).

6.2.3 Population structure methods

Three different methods are presented here to analyze population structure in the 38 ash trees; STRUCTURE, plastid haplotype networks, and PCA. The inputs for STRUCTURE analysis and PCA originate from the same data; the set of 394,885 SNP loci with genotype information for every sample. The data for the plastid haplotype network are sequence alignments of the long single copy region of the plastid chromosome, which clusters sequences based on the number of base substitutions between them.

STRUCTURE v2.3.4 was used to cluster the 38 trees into a pre-defined number of 'k'

groups. As the input data for STRUCTURE should be a list of unlinked loci with genotype information for all samples, starting from my set of 394,885 SNP loci, I first extracted all scaffolds with ten or more SNPs; these totalled 8,955 scaffolds. I then selected three random SNPs on these scaffolds and piped them into three separate STRUCTURE input files. Therefore I had three files each with 8,955 different SNPs, making a robust set of three independent replicates. I ran each of the three files through k=1 to k=20 (one rep for each file, three reps in total), with 100,000 burn-in, 100,000 reps, with admixture assumed and no prior population information given. Results were run through StructureHarvester Web v0.6.94 [Earl & VonHoldt 2012] to calculate the Delta K; a measure of how well each value of k maximises the likelihood of the result given the starting data [Evanno et al. 2005]. K=3 was found to have the highest delta k value, therefore the results for k=3 were then run through CLUMPP v1.1.2 to sort the order and group membership for the three runs of STRUCTURE. Re-ordered results were then imported back into STRUCTURE v2.3.4 to generate Q-value bar plots. These steps allow the plots of the three runs of STRUCTURE to have the same sample order so that the plots can be stacked and visualised together.

To analyze relationships among plastid sequences of the 38 trees, consensus sequences of the large single copy region (LSCR) of the plastid genome were extracted for each of the 38 samples from their read mappings, using the 'Extract Consensus Sequence' tool in the CLC Genomics Workbench v8.0. The sequences were aligned using the 'Create Alignment' tool, and the alignment exported in Phylip format. This was then imported into PopArt v1.7 [http://popart.otago.ac.nz], where a Median-Joining network was generated.

PCA was used as a robust way of clustering the 38 trees into groups, without assuming any particular features about the data or samples. PCA reduces the complexity of multifactorial data by describing most of the variance of the data in the first few Principal Components (PCs). The initial data input was again the set of 394,885 SNP loci with genotype information for every sample. All PCA steps were carried out using the SNPRelate v1.2.0 Bioconductor R package. First, the VCF file of variant information was converted to GDS using the command 'snpgdsVCF2GDS', and the GDS file was loaded using 'snpgdsOpen'. The command 'snpgdsLDpruning' with the threshold of 0.1 was used to remove SNP loci that give mostly the same information due to being in high linkage disequilibrium. This reduces the size of the dataset and allows the program to run much faster. After the LD pruning step, 34,607 SNP loci remained. The command 'snpgdsPCA' was then run with the reduced set of SNPs to produce the PCA results, which were plotted within R v2.15.2.

6.2.4 Effective population size methods

The PSMC method uses the distribution of TMRCA between alleles in a single genome to estimate the change in N_e over time. The program is available for download from https://github.com/lh3/psmc. I conducted two analyses with PSMC v0.6.5; one on the British reference tree, and one on Danish Tree35, found to have low susceptibility to ADB [McKinney et al. 2011]. Reads for Tree35 were downloaded from the Open Ash Dieback ftp site as described in Section 6.2.1. I selected libraries from Tree35 that would give a similar coverage to the combined mapping of all the short-insert reads of the British tree

(200, 300 and 500 bp libraries). The starting input data for PSMC in both cases was a .bam alignment of reads, obtained by using the 'Map Reads to Reference' tool within the CLC Genomics Workbench v8.0, with length fraction set to 0.95 and similarity fraction to 0.9. The mapping was exported in .bam format, and a consensus sequence was obtained following PSMC recommendations, by using samtools v0.1.18 mpileup command with options: -C 50 -A -Q 20 -u, bcftools v1.1 to convert the bcf file to vcf format, and finally using vcfutils.pl to convert the vcf file to a consensus sequence where the coverage was between 5 and 200. The PSMC program was then run with default parameters except for: -p 4+25*2+4+6. It also provides options for running bootstraps by splitting the genome into segments (I chose 100), using the 'splitfa' command, and running the PSMC program on each segment. Generation time was set to 15 years (approximate age to reproduction in ash), so that plots used time in years rather than number of generations, and mutation rate used was $7.5e^{-9}$ (based on mutation rate measured in *Arabidopsis thaliana* in Beilstein et al. (2010)). Results were then plotted in R v 2.15.2.

A complementary method was used to estimate N_e, as the inference from PSMC is limited to certain time intervals no more recent than 200,000 years ago. Whereas, LD provides estimates of very recent N_e. The program SNeP v1.1 [Barbato et al. 2015] was used to calculate N_e from measures of LD using genome-wide polymorphism data from a population of samples. For this I used the set of 394,885 SNP loci from the 38 trees described in section 6.2.1. The VCF file of the 394,885 SNPs was converted into .map and .ped files suitable as input for SNeP. The third column in the Map file (linkage distance in Morgans) was set to zero for all SNPs, as these values were unknown and SNeP calculates this value from each SNP's physical distance anyway. SNeP was then run with a minimum distance between SNPs of 10,000 bp and a maximum of 400,000 bp, with Sved's modifier for recombination rate, and with 50 bins. All other parameters were kept as default. Estimated effective population sizes over time, as well as LD decay over genomic distance, were plotted in R.

6.3 Results and Discussion

6.3.1 Analysis of population structure

STRUCTURE

STRUCTURE v2.3.4 was run in three independent runs each of 8,955 SNPs from all 38 trees, for k=1 through to k=20. The Delta k plot for k=1 to k=20 shows that k=3 had the highest value (Fig. 6.5), and is therefore the most likely number of clusters. The results for three runs of STRUCTURE at k=3 are shown in Fig. 6.6 and geographically in Fig. 6.7.

The first noticeable pattern, is the stark difference of two trees, 37 and 6, compared to the rest. These two trees consistently show the vast majority of their Q-value belonging to the green group across all three replicates. Very few of the remaining trees show any membership at all with this group, and instead show varying levels of membership to the red and blue groups. In addition, when STRUCTURE was run with k=2, all other samples

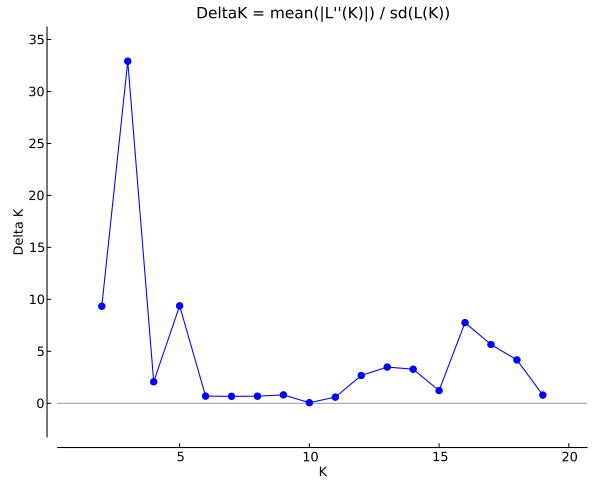


Figure 6.5: Plot of Delta K values from k=2 to k=19 (the two extreme values of k, 1 and 20, are not shown in plot).

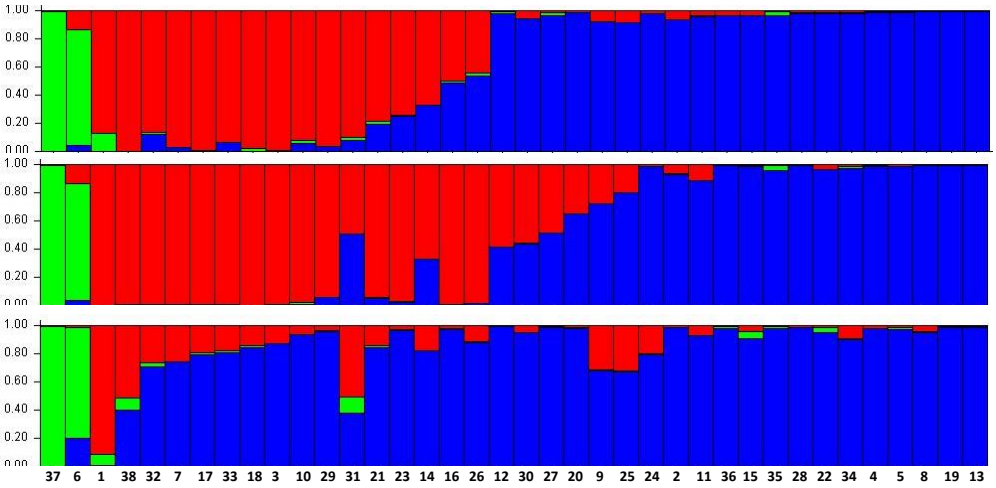


Figure 6.6: Three independent runs of STRUCTURE at k=3, each using 8,955 SNPs.

were shown as belonging to the same group (as opposed to trees 37 and 6 being merged with another group), showing that these two samples are really quite different to the others. When viewed on the map, the two 'green' samples are not located particularly close to each other; sample 37 being in northern Germany and 6 being in western Belgium. One hypothesis for why these two trees are outliers, is that they could belong to a genotype of *F. excelsior* that we undersampled by chance. Further sampling in Belgium, Netherlands and Germany could confirm whether any more trees belonging to the green group can be found. Another hypothesis is that the two trees could be hybrids, or descended from hybrids, of *F. excelsior x angustifolia*. This is entirely plausible since previous studies have found that hybrids readily occur in hybrid zones [Heuertz et al. 2006, Gerard et al. 2013], and although the main range of *F. angustifolia* does not reach to Belgium or Germany, there are patches of *F. angustifolia* found throughout Germany (Fig. 6.2). Therefore these two trees could potentially coincide with small pockets of *F. angustifolia*. Another explanation could be that the ancestors of these trees have carried *F. angustifolia* alleles from hybridisation during glacial periods in refugia, as they re-colonised northwards in Europe. Although the *F. angustifolia* loci would get purged from the genome over successive generations with-

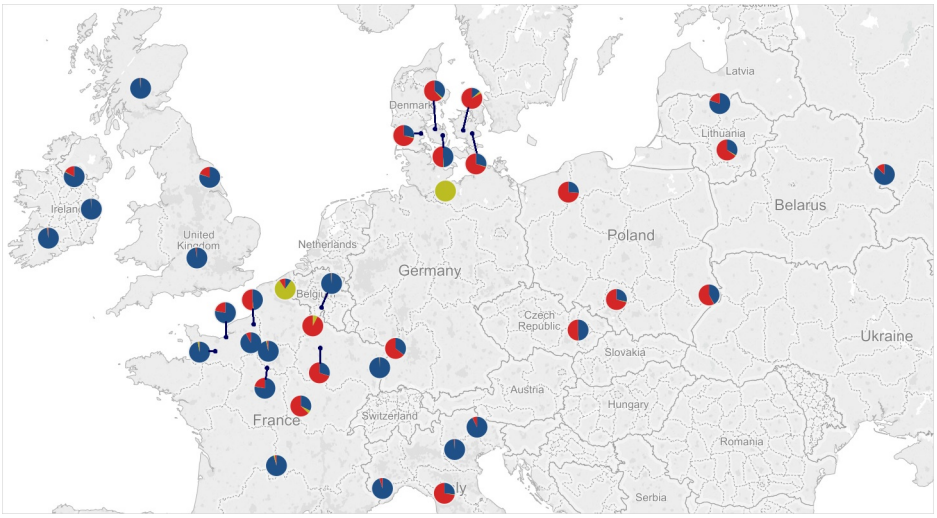


Figure 6.7: Map showing STRUCTURE results for 37 trees. Size of pie slices represent average membership of sample to each coloured group, calculated from results of Fig. 6.6. Sample 34 is left off map due to uncertain source location.

out continuous hybridisation, perhaps trees 6 and 37 are carrying a sufficient amount of *F. angustifolia* remnants to make them different from the other 36 trees that, I assume, are "pure" *F. excelsior*. Further genetic testing of these two trees could shed some light on their genetic background, for example, by using markers that have already been used to distinguish the two species [Gerard et al. 2013; Thomasset et al. 2011; Heuertz et al. 2006]. As mentioned in the "Current Population Research" section, possible movement by humans should be taken into account here, when two outliers appear as in this case. Although the source populations are all natural forests and not plantations stocked from elsewhere, perhaps areas nearby to these two trees were stocked with ash from southern or central Europe where a high frequency of hybrids exist. As gene flow is very high in this wind-pollinated outbreeding species, hybrid DNA could easily pass from re-stocked plantations into natural populations.

There also does not appear to be a strong pattern regarding the red and blue groups. However, it is clear that all the samples from the UK are mostly blue, and the red group tends to appear in central and northern European trees. This seems to fit with previous findings using nuclear microsatellites [Huertz et al. 2004a], which found a mixture of genotypes in central and western Europe with some distinct genotypes in south-eastern Europe, which was unfortunately outside the range of the 38 trees I have used. It is also worth noting that many of the studies previously reporting on population structure in ash have used many individuals per population site, while I have only one individual per location, albeit with many more genomic markers. Still, additional sampling would certainly help elucidate some unclear patterns here.

Plastid haplotype networks

To analyze relationships among plastid sequences of the 38 trees, consensus sequences of the large single copy region (LSCR) for each of the 38 trees were aligned. A Median-Joining network was generated from the haplotypes (Fig. 6.8). All trees within each coloured group are separated by no more than five base substitutions. Groups shown in coloured boxes in Fig. 6.8 are shown on a map in Fig. 6.9.

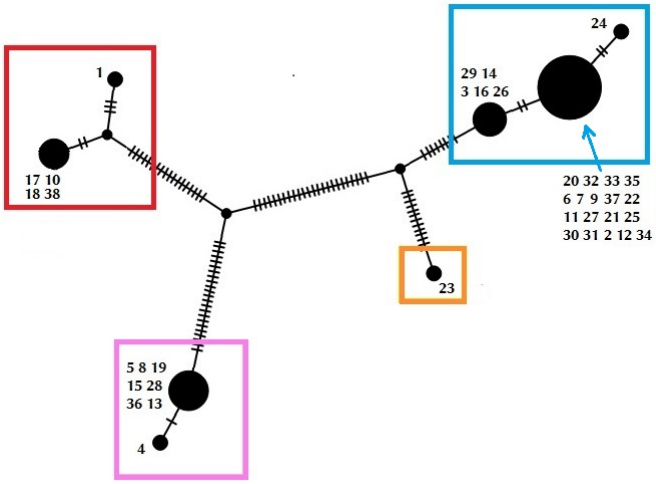


Figure 6.8: Median-joining network of plastid (long single copy region) haplotype sequence alignment. Each notch represents one base substitution. All trees within each coloured group are separated by no more than five base substitutions.

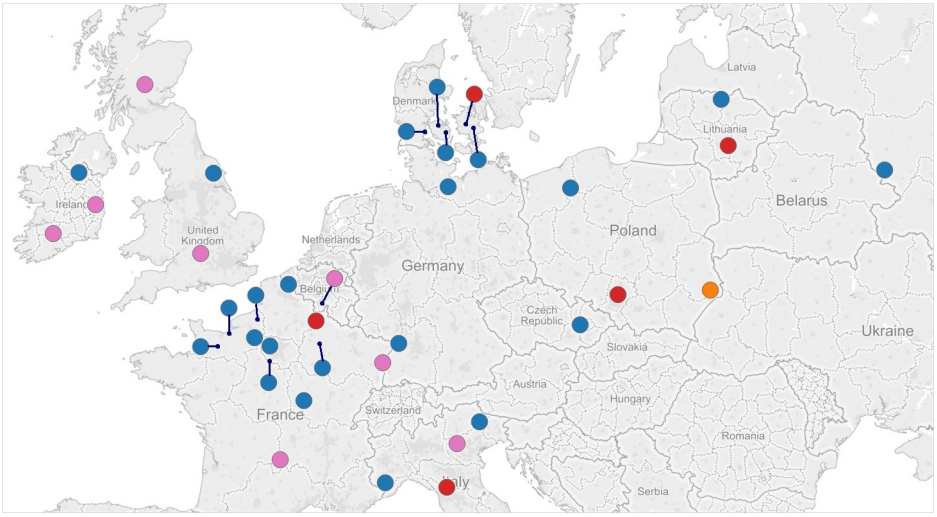


Figure 6.9: Map of ash diversity samples coloured by plastid haplotype as shown in Fig. 6.8. Sample 34 left off map due to uncertainty of source location.

There appears to be no definite geographical pattern for these groups. Trees in the pink group *tend* to be located more to the west as there are no pink samples east of Italy, and trees in the red group *tend* to be located more to the east as there are no red samples in the UK. However it is not a clear pattern and perhaps additional sampling would reveal an equal spread of all groups. Interestingly, the two outlying trees from the STRUCTURE analysis, 6 and 37, have identical LSCR sequences to many other trees, and are therefore definitely not outliers in this analysis.

Principal Components Analysis

Principal Components Analysis was performed on an unlinked set of 34,607 loci in the 38 trees. The first four PCs were plotted against each other, but only PC1 Vs PC2 and PC2 Vs PC3 produced results that showed any kind of pattern. Remarkably, PC1 Vs PC2 (Fig. 6.10 shows clusters of trees very similar to those found by STRUCTURE, and PC2 Vs PC3 6.11 shows clusters very similar to the groups found by the plastid haplotype network. These PCA results lend some support to the previous STRUCTURE and plastid haplotype results, as in both cases, two different methods have found very similar results.

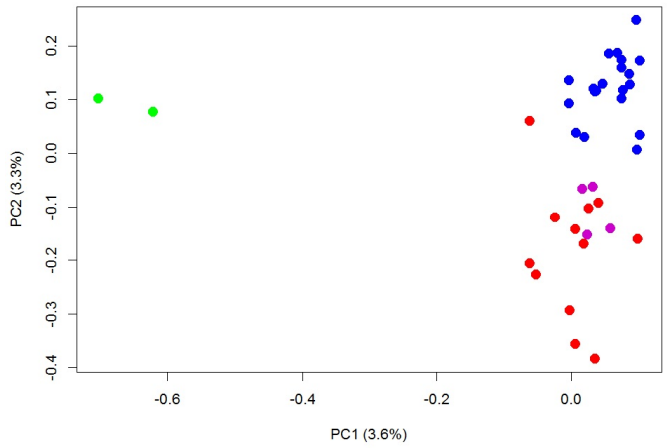


Figure 6.10: PC1 vs PC2 with points coloured by the STRUCTURE group to which the sample has the largest average membership, as shown in Fig. 6.6. Four trees (14, 16, 23 and 26) were almost equally assigned to red and blue groups on average, therefore they are coloured purple.

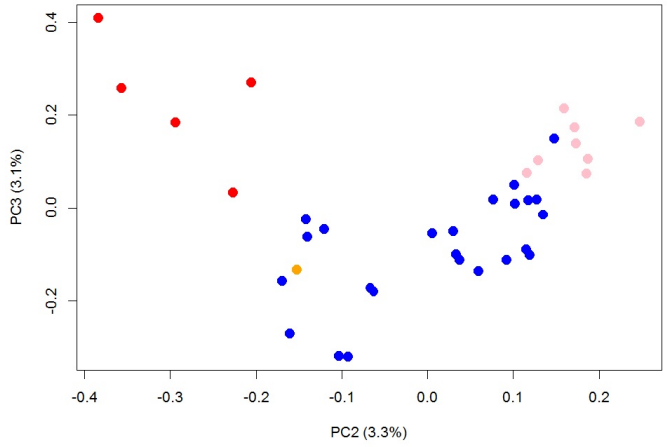


Figure 6.11: PC2 vs PC3 with points coloured by plastid haplotype network groupings, as shown in Fig. 6.8.

6.3.2 Estimating effective population size history

PSMC estimates past N_e

Fig. 6.12 shows that the N_e for both Tree35 and the British tree (referred to as 'BATG' in the figure), increases and peaks at approximately 20 mya. Since then, the N_e for both trees has decreased substantially until the most recent point that PSMC can detect, approximately 200,000 years ago.

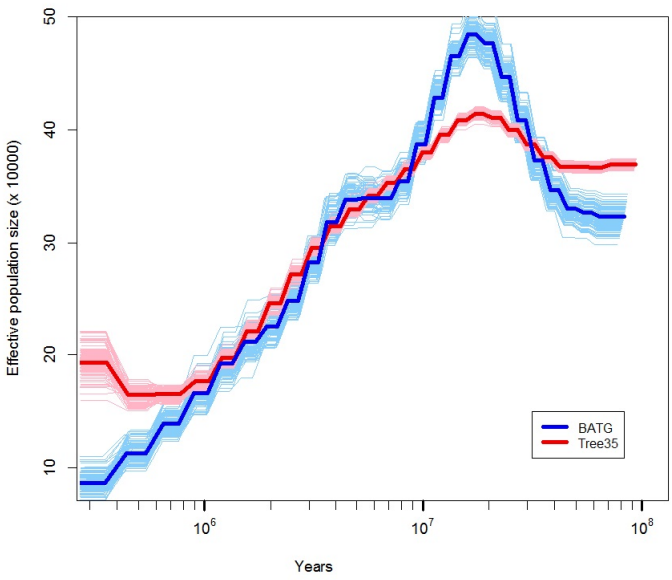


Figure 6.12: Results of PSMC analysis show an initial increase in N_e until approx. 20 mya, and a subsequence decrease for both the British ash tree and Tree35. N_e for Tree35 slightly increases between 400 and 200 kya, while the British tree's N_e estimate continues to fall.

This could reflect the overall cooling of the planet from the warm Oligocene epoch (ended 23 mya) through the Miocene (23 - 5 mya) and to the cool Pliocene epoch (5 - 2.5 mya). A major ice age started at the end of the Pliocene on the boundary with the Quaternary period, which itself has been subjected to repeated glaciation events. Generally, glaciation periods caused the ranges of many European species to shift southwards to reside in refugia until the climate warmed enough for them to colonise northwards again. Many studies have found evidence for refugia in southern and eastern Europe, such as Italy, Iberia, and the Balkan peninsula [Heuertz et al. 2004a; Heuertz et al. 2004b; Magyari et al. 2014]. This is likely the case with the European ash; a genetic bottleneck caused by the glaciation periods would reduce the effective population size.

Interestingly, the N_e for Tree35 starts to increase slightly between 400,000 and 200,000 years ago, while the British tree's N_e continues to decrease. It could be that the ancestor population for Tree35 started to expand around this time, leading to an increase in N_e. However, the differences in genomic diversity could also possible cause this. The British tree is the progeny of a self-pollination and therefore its genomic diversity will be less than the true diploid Tree35. A reduction in heterozygosity will almost certainly have an impact on a measure of N_e that is calculated from the TMRCA of alleles in the genome.

In hindsight, the mutation rate used for this analysis ($7.5e^{-9}$) based on a reported mutation rate in *Arabidopsis* was probably not appropriate for a long-lived tree species with a much larger generation time such as ash. In addition, upon re-reading the documentation for the PSMC program, the mutation rate should be per generation and not per year. In the absence of a known mutation rate in ash, using a mutation rate for a tree species with similar life history traits would have been more appropriate, such as that of poplar ($2.84e^{-9}$ per year as reported in Busciazzo et al. (2012)). Therefore the mutation rate per generation (assuming a generation time of ~15 years) would be approximately $4.2e^{-8}$. This would not

have changed the overall pattern of N_e history in Fig. 6.12 however, only the timescale; using the revised mutation rate would have shifted the whole graph in Fig. 6.12 to the left. However, without an estimated mutation rate for ash itself we cannot be sure that the revised graph would be a good indication either. Therefore, we will have to settle with the assumption that the general pattern showing decline in N_e over time is correct, but that we cannot be sure of the timescale for this decline.

Linkage disequilibrium estimates recent N_e

Fig. 6.13 shows continued decrease in N_e between approximately 4000 and 100 generations ago (around 60,000 to 1500 years ago, assuming a generation time of 15 years). Similar to the analysis with PSMC, a continuing decrease could be due to repeated glacial cycles of the late Pleistocene until approximately 11,000 years ago. However, the estimated N_e values do seem very low. Barbato et al. (2015) explain possible inaccuracies in the method at the most recent and oldest N_e estimates: "for recent generations, large values of c are involved, not fitting the theoretical implications that Hayes proposed to estimate a variable N_e overtime (Hayes et al.,2003). Estimates for the oldest generations might also be unreliable as coalescent theory shows that no SNP can be reliably sampled after $4N_e$ generations in the past (Corbin et al., 2012). Further, N_e estimates, and especially those related to generations further in the past, are strongly affected by data manipulation factors, such as the choice of MAF (Minimum Allele Frequency) and alpha values. Additionally, the binning strategy applied can interfere with the general precision of the method, for example where an insufficient number of pairwise comparisons are used to populate each bin." Therefore, perhaps the values of estimated N_e cannot be completely trusted. However, the general trend of decrease over time is likely to be correct.

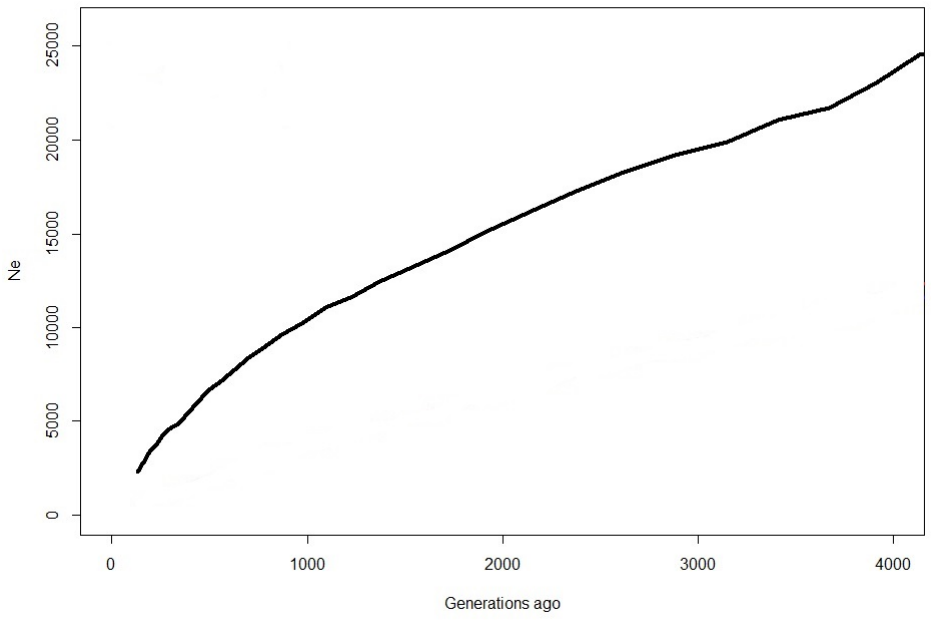


Figure 6.13: Results from SNeP show that N_e has decreased over the past 4,000 generation

Fig. 6.14 shows that even between SNPs just a few hundred bases apart, the average LD is 0.15 which is already fairly low. Typically forest trees have been shown to have very low LD due to an outcrossing mating system [Ingvarsson et al. 2016]. Some older studies on trees have shown LD decaying from very high (0.4-0.5) to low levels within a few hundred or thousand bases, e.g., *Eucalyptus* [Thavamanikumar et al. 2011], *Pinus taeda* [Brown et al. 2004]. A summary table of LD values in various tree species is shown in Table 6.15. However, these studies are limited to approximately 5kbp between loci, as they all tend to study specific sets of genes rather than approaching LD on a whole-genome scale. There are also some other outcrossing tree species showing much higher LD than ash at large distances, such as *Prunus avium* which has $r^2 = 0.2$ at 100kbp [Campoy et al. 2016]. One main difference in the three aforementioned studies is that LD was estimated from very few SNPs; in the region of a few hundred or thousand at most, from selected genes or loci of interest. This limits our ability to compare these LD estimates to ash.

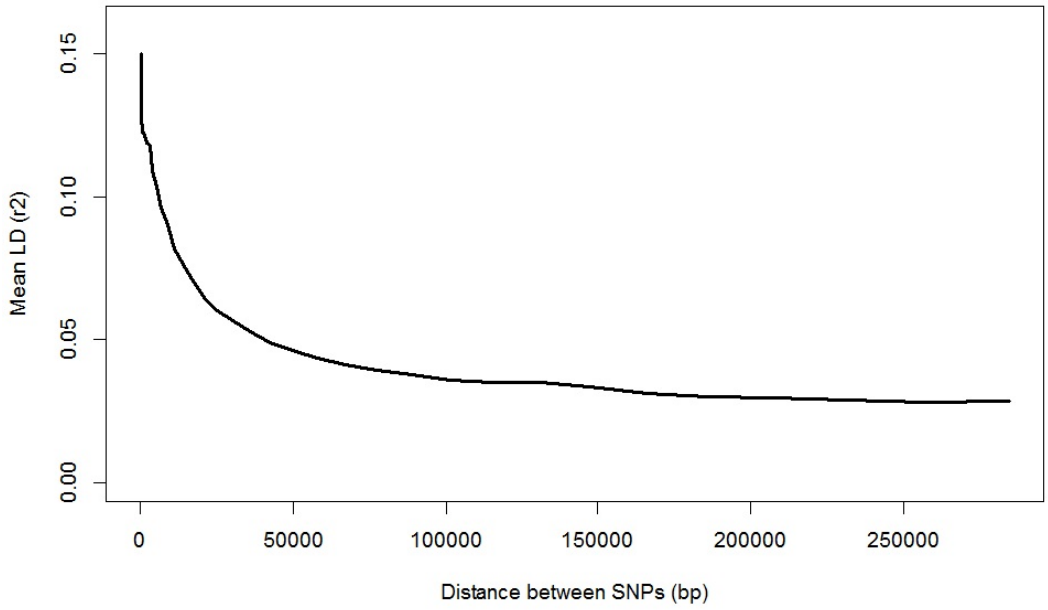


Figure 6.14: Linkage disequilibrium decay over increasing distance between loci

In this study, we use more than 300,000 genome-wide SNPs, without any preference for those residing in a gene region or any other particular location. Therefore we can only compare the LD estimates for ash with studies that use similar whole-genome methods. More recent studies on LD in forest trees using comparable whole genome methods, have found that LD decay is in fact slower than expected from previous single-gene studies, and is a lot more variable [Ingvarsson et al. 2016]. For example, a recent study on three species of *Populus* showed very similar mean LD values to ash, over a similar range of SNP distances using between 710,000 and 1.4 millions SNPs [Wang et al. 2016]. In fact, the LD decay curve for *Populus tremuloides*, Fig. 6.16 is almost identical to that of ash, e.g. approximately r^2 of 0.1 at 10kbp and eventually plateauing at around 0.05. However, the r^2 estimates of *P. tremuloides* at very small distances are much higher than that of ash, likely

Common name	Scientific name	Function of genes studied	Number of genes studied	Mean total nucleotide diversity (θ_π)	LD decay	References
Loblolly pine	*Pinus taeda*	Wood formation	19	0.00398	Within 2,000 bp	Brown et al. (2004)
		Drought response	18	0.00507	Within 800 bp	Gonzalez-Martınez et al. (2006)
Maritime pine	*Pinus pinaster*	Wood formation	8	0.00241	–	Pot et al. (2005)
		Drought response	11	0.00548	–	Eveno et al. (2008)
Radiata pine	*Pinus radiata*	Wood formation	8	0.00186	–	Pot et al. (2005)
Scots pine	*Pinus sylvestris*	Cold-related	14	0.00600	Within 200 bp	Wachowiak et al. (2009)
		Allozyme coding	6	0.01000[a]	No appreciable decay	Pyhajarvi et al. (2011)
Douglas Fir	*Psuedotsuga menziesii*	Cold Hardiness and wood formation	18	0.00655	Within 1,500 bp	Krutovsky and Neale (2005)
		Cold Hardiness	121	0.00435	Within 1,000 bp	Eckert et al. (2009b)
Norway spruce	*Picea abies*	Growth cessation	22	0.00208	Within 100 bp	Heuertz et al. (2006)
White spruce	*Picea glauca*	Multiple functions	105	0.00430	Within 65 bp	Pavy et al. (2012)
European aspen	*Populus tremula*	Multiple functions	5	0.01110	Within 500 bp	Ingvarsson (2005)
		Multiple functions	77	0.00420	Within 200 bp	Ingvarsson (2008)
Balsam poplar	*Populus balsamifera*	Multiple functions	590	0.00280	No appreciable decay	Olson et al. (2010)
Japanese cedar	*Cryptomeria japonica*	Multiple functions	7	0.00251	–	Kado et al. (2003)
		Multiple functions	5	0.00213	–	Kado et al. (2006)
		Multiple functions	10	0.00156	–	Kado et al. (2008)

Figure 6.15: Additional comparison of tree LD values, from studies on particular genes. Table taken from Thavamanikumar et al. (2013); references contained therein.

reflecting a higher resolution of the method used in Wang et al. (2016). The researchers demonstrate that LD is indeed incredibly variable over physical distance between SNPs, with some markers over 10kbp being in near complete LD. Similarly, an earlier study of LD in *Populus trichocarpa* [Slavov et al. 2012] found a slow decay of LD, with the r^2 only dropping below 0.2 at 6-7 kbp between SNPs. A study of LD in *Eucalyptus grandis* using 21,000 SNPs also suggested a slow decay in LD (with r^2 of 0.1 even at 40 kbp), and again that LD is variable, with some SNPs 50kbp apart still being in near complete LD [Silva-Junior & Grattapaglia 2015]. On the other hand, a recent study of LD in loblolly pine [Lu et al. 2016] shows very rapid decay in LD, with an r^2 of 0.05 already at 450 bp. Though, the results of LD in a gymnosperm with a very large genome may not be very comparable to that of ash or indeed other angiosperm trees. In addition, the results of Lu et al. (2016) were generated from exome rather than whole-genome sequencing. It is clear that patterns of LD decay differ between tree species with different genome sizes, population structure, life history traits and breeding system, but also that whole-genome methods can paint a much bigger picture then what was previously available using markers from only a few select genes.

6.4 Conclusion and future directions

Previous research has shown a variety of patterns in the population structure of *F. excelsior* in Europe, differing between the types of markers used. Chloroplast DNA markers have tended to find highly structured and differentiated populations, with genotypes clustered in specific regions [Heuertz et al. 2004b; Tollesfrud et al. 2016]. In contrast, studies using

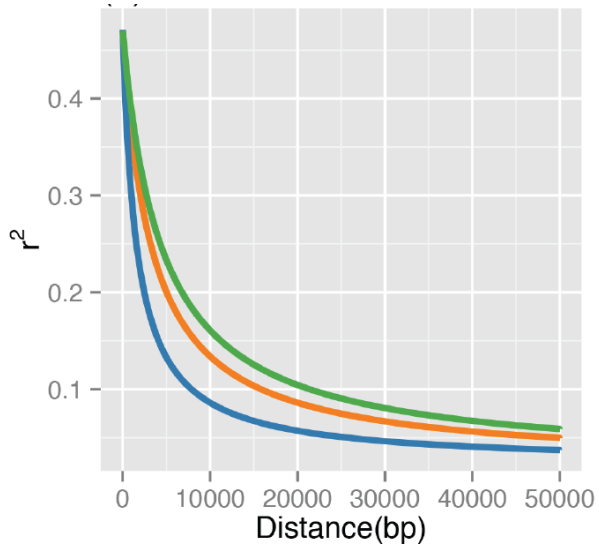


Figure 6.16: Mean LD decay with physical distance in three *Populus* species, *P. tremula* (orange line), *P. tremuloides* (blue line), and *P. trichcarpa* (green line). Image taken from Wang et al. (2016).

nuclear markers have found a largely homogenous population across western and central Europe with separate demes in the east and in Sweden [Heuertz et al. 2004a; Sutherland et al. 2010; Beatty et al. 2014; Tollesfrud et al. 2016]. Much research has identified glacial refugia in the Carpathian mountains, and Iberian and Balkan peninsulas [Magyari et al. 2013; Heuertz et al. 2004b; Heuertz et al. 2004a], which is also the case for many other European trees. Together, the evidence of refuge locations and current genotype distributions has led to suggestions of various post-glacial colonisation routes; from refugia in Italy to France and southern Germany, from Iberia to the UK, from the eastern Alps to Germany, Denmark, the Czech Republic and Poland, and from the Balkan Peninsula north into the Baltic states and northern Scandinavia [Heuertz et al. 2004b; Magyari et al. 2013; Tollesfrud et al. 2016].

Using nearly 400,000 genome-wide markers, I find that *Fraxinus excelsior* appears to have very little population structure across Europe, consistent with other studies of ash populations in central Europe using nuclear markers showing a large but diverse population [Heuertz et al. 2004a; Tollesfrud et al. 2016]. Unfortunately we did not have access to any samples from the locations previously suggested as glacial refugia, which could have shown high differentiation compared to the rest. Perhaps a clearer pattern of population structure would have been found if these regions had been included; i.e. the refugia trees acting as an outgroup if they have sufficiently diversified genomes. The remainder of samples from central Europe could have then been classed into a single diverse population if they are all more similar to each other than to trees from the refuge location. On the other hand, the similarity of the trees to their various refuge locations (from predicted colonisation routes) could have become apparent, and geographical patterns may have emerged. Nevertheleess some various groups of genotypes were identified by different methods, such as STRUCTURE, PCA and plastid haplotype networks, but none show a clear geographical pattern. Two trees however showed very different genotypes compared to the rest in the STRUCTURE analysis, as they were placed very distinctively in their own group with little membership to either of the other two groups. I hypothesise that these two (#6 and #37) could be hybrids, or descended from hybrids, of *F. excelsior* and *F. angustifolia* (narrow-leaved ash), another European species whose range overlaps with that of *F. excelsior* [Gerard et al. 2013; Heuertz

et al. 2006].

By studying the effective population size history using the reference genome, Danish Tree35 and the 38 Diversity panel trees, I find that effective population size has been decreasing since approximately 20 mya. Considering trends in climate at the time, the decrease is likely caused by global cooling which restricted the ranges of many plants and animals into refugia. Although we cannot be completely certain of the timings due to lack of knowledge of the mutation rate in ash, we can be more certain of the general pattern showing the decrease in effecive population size over the last 10 million or so years. Even in the recent past (since the LGM and expected range expansion), the effective population size of ash is still showing a decline. This has been found in other studies that suggest that ash was harvested intensely during the bronze age (around 5000 years ago) for wood and fodder [Thomas 2016].

The wider implications of these findings are that, firstly, breeding programs for ash (in response to the population decline caused by ash dieback) should aim to maintain as much of the genetic diversity of the European population as possible. Since the effective population size of ash has been constantly decreasing, and will continue to decrease in the coming years due to ADB and climate change [Goberville et al. 2016], ash is likely going through a bottleneck stage and therefore the gene pool may be severely restricted. It is important to maintain genetic diversity so that the population has variation to adapt to future environmental or biological stresses. Region-specific genotypes are apparently only found within the plastid chromosome [Heuertz et al. 2004b; Tollesfrud et al. 2016], therefore diversity markers used in breeding programs should be enriched for plastid DNA if these genotypes are to be preserved.

Future studies on population structure in ash could extend our study by sampling additional locations, and by sampling more individuals in each location (thus providing a more commonly used input dataset for analyses such as STRUCTURE). We were very fortunate to be able to use a trial of ash trees growing in the UK that originate from around Europe, which allowed us rapid access to range-wide genetic material. In light of the recent spread of ADB, quick action was needed to gain as much information on the population structure of European ash as possible. However a more in-depth study with more samples, especially more in Eastern Europe (where our sampling was sparse) could shed light on the various clusters of genotypes that I find and whether these are also present in Eastern European trees. Another research question could be whether the genotypes identified in specific trees are also found in neighbouring trees. This could elucidate whether certain genotypes tend to dominate localised populations, or whether they are actually more diverse and each tree can be very different to its neighbour. In the latter case, it could be more correct to think of the whole European ash population as one single very diverse population. In addition, as previously mentioned we did not obtain any samples from the glacial refuge locations. Samples from these areas would provide an interesting genetic comparison with those already obtained, as additional, diverse genotypes could still remain in these locations, as found by previous studies [Heuertz et al. 2004a; Heuertz et al. 2004b; Magyari et al. 2013].

Further research into the two very different trees (#6 and #37) could also be carried out

to elucidate whether these are in fact hybrids as suspected, or actually a very rare *F. excelsior* genotype. This need not use whole genome sequencing, and could make use of some of the polymorphic loci identified in this study in order to genotype several *F. angustifolia* trees and compare the results with trees #6 and #37. In addition, several microsatellite markers for the two species have been used successfully in other studies to differentiate the species [Thomasset et al. 2011; Thomasset et al. 2013; Gerard et al. 2013]. If they do indeed possess high similarity to *F. angustifolia* genotypes, this would add to the knowledge of known natural hybrid locations as have already been found in the range overlap areas in France, Spain and Italy [Gerard et al. 2013], and could also lend support to theories on the post-glacial range expansion of both species [Heuertz et al. 2006].

With climate change predicted to alter the habitable zones of many temperate species, an interesting long-term study would be on the range shift of *F. excelsior* over the course of the next 20 or so years. Coupled with the host-pathogen dynamics with *H. fraxineus*, and the effect of climate change on both species, the range shift of *F. excelsior* could follow several different scenarios. Goberville et al. (2016) modelled these scenarios with various climate predictions. With increased warming, the authors predict that ranges of both species separately to move north- and eastward and to regions with high altitude. However when the interaction between the two species is considered, the range of *F. excelsior* moves away from northern and central Europe due to the presence of the pathogen in these areas. An aspect not considered in this paper however are the potential effects on the genetic pool of *F. excelsior*. A restriction of populations in central and northern Europe may cause a loss of certain haplotypes that are only found in those areas, particularly chloroplast haplotypes that have been shown to be very differentiated across Europe. In addition, the remaining ash populations may show evidence of a genetic bottleneck with a range restriction and possible purifying selection for traits such as heat tolerance or low susceptibility, the latter of these having only a very narrow prevalence in the natural population currently [McKinney et al. 2011; Harper et al. 2016; Sollars et al. 2017].

Chapter 7

Epigenetic variation in isogenic samples

7.1 Introduction

7.1.1 DNA methylation in plants

There has been much evidence that epigenetic changes in plant genomes can alter phenotypes by regulating gene expression. Epimutations occur much more frequently than genetic mutations (e.g. in *Arabidopsis*: c. 4.5×10^{-4} epimutations per CG site per generation [Schmitz et al. 2011] versus c. 7×10^{-9} base substitutions per generation [Ossowski et al. 2010]) and are heritable both mitotically (through cell replication) and meiotically (to gametes and therefore the next generation). There is an increasing number of plant traits that have been found to be under epigenetic control [Niederhuth & Schmitz 2014], such as fruit ripening in tomato [Manning et al. 2006] and energy use efficiency in canola [Hauben et al. 2009]. Therefore epigenetics is of much interest to plant breeders [Springer 2013; Bilichak & Kovalchuk 2016] as a source of trait variation in addition to genetic factors and especially due to the heritability of some epigenetic marks, which renders them open to artificial selection [Hauben et al. 2009].

Methylation of cytosine is one such epigenetic modification that can alter gene expression. In contrast to mammalian genomes, cytosines in plant genomes can be methylated in all three sequence contexts: CG dinucleotides are the most frequently methylated, while CHG and CHH (where H = A, C or T) are less commonly methylated. The methyl groups are added and maintained over cell generations by methyltransferase enzymes (Law & Jacobsen 2010). DRM2 catalyzes de novo methylation in all three contexts, mediated by small interfering RNAs of the RNA-directed DNA Methylation pathway, while methylation is maintained in the CG context by MET1 and in the CHG context by CMT3 [Zhang & Zhu 2011].

Methylation serves various purposes within plant genomes via regulating gene expression. Polyploidy is a common phenomenon in the plant kingdom, with frequent Whole Genome Duplication (WGD) events identified throughout numerous lineages [Blanc & Wolfe 2004; Jiao et al. 2011]. As the amount of genetic material present is doubled upon polyploidisation, silencing pathways such as RdDM are commonly employed to compensate for the increased gene dosage (Schmitz et al. 2013a; Sehrish et al. 2014). Similarly, genomic im-

printing involves the silencing of either the maternal or paternal copy of a gene so that mRNA is transcribed from only one copy. Finally, the vast majority of transposable elements are highly methylated [Becker et al. 2011; Schmitz et al. 2013b] in order to silence their activity. Transposons insert themselves into random places in the genome, possibly inside genes, therefore methylation serves as a genome defence mechanism to shut down their expression [Michael 2014].

Although DNA methylation has been studied in plants for many years, the vast majority of focus has been on crop plants (e.g. rice [Li et al. 2012], maize [Wang et al. 2015b; Li et al. 2014a], soybean [Schmitz et al. 2013a]) and model organisms such as *Arabidopsis* [Becker et al. 2011; Schmitz et al. 2011]. To my knowledge, only three tree species have had whole genome bisulphite sequences published; model tree species *Populus trichocarpa* [Liang et al. 2014], Norway spruce (Picea abies)[Ausin et al. 2016], and *Betula platyphylla*, (white birch, [Su et al. 2014]). In addition, oil palm *Elaeis guineensis* has had whole-genome bisulphite sequencing performed, but this has only been described in very little detail in Ong-Abdullah et al. (2015), and a project using bisulphite sequencing on *Populus tremula x alba* is detailed at http://genome.jgi.doe.gov/PoptreSeqsample8. The methylomes of some trees such as oak [Platt et al. 2015; Gugger et al. 2016] have been sequenced using Reduced Representation Bisulphite Sequencing (RRBS), which sequences the regions surrounding restriction sites. This technology allows cheaper sequencing with deep coverage of many loci, but is not completely genome-wide. In addition, a study on *Populus deltoides* leaf methylomes uses MEDIP-seq (Methylated DNA immunoprecipitation sequencing) [Gao et al. 2014].

7.1.2 Background to the ash methylome project

As described in previous chapters, European ash (*Fraxinus excelsior*) has been the subject of a large genome sequencing project in response to the rapid spread of ash dieback disease in Europe. One important finding from the project is the recent WGD event shared with *Olea europaea* and therefore thought to be common to the Oleaeceae family (see Chapter 5). In addition, a less recent WGD was identified, shared with olive and possibly the other Lamiales species *Mimulus guttatus* and *Utricularia gibba*. Other research has demonstrated (i) extensive epigenetic reprogramming after a WGD or polyploidisation event (e.g., Levy & Feldman 2004; Salmon et al. 2005; Chen 2007; Ksiazczyk et al. 2011; Bao & Xu 2015), and (ii) that many paralogs retained after WGD events are unequally silenced via DNA methylation or by other means, as a way of reducing the doubled gene expression back to normal levels [e.g. Schmitz et al. 2013a; Sehrish et al. 2014; Wang et al. 2015a]. I therefore hypothesise that many of the homeologs retained in the ash genome since the WGD events may be unequally silenced via differential methylation as a means of gene dosage compensation. I investigate this by comparing the methylation levels in each member of an homeolog pair, and test how many homeolog pairs have significantly different methylation levels.

Another important finding from our collaborators was the identification of a number of genes whose expression was significantly associated with susceptibility to ADB. I investigate the methylation level in these genes to see whether the expression difference could be caused

by varying methylation levels. I also attempt to identify differentially methylated regions (DMRs) that could be associated with ADB susceptibility. However, due to a low number of samples available (see Section 7.2.1), I stress that these results are very preliminary, due to the likelihood of methylation differences in a small group of samples being due to chance, as well as the low power of rejecting the null hypothesis, especially with small effect sizes.

Research on other plants has shown that DNA methylation does have a role to play in resistance (for example, demethylation of a promoter region enables expression of a resistance gene in rice [Akimoto et al. 2007]) and in response to infection (for example, demethylation in response to infection in rice [Sha et al. 2005]). Notably, Verhoeven et al. (2010) found that pathogen and herbivore stress induce varying epigenetic changes in dandelions, meaning that the plants responded to infection in different ways. This induced epigenetic variation could therefore explain the varying ADB susceptibility phenotypes present in ash, which previous research has shown are not discrete categories but actually a continuous distribution of damage scores.

In this chapter, I present the methylome of *F. excelsior*, which has not been sequenced before. I describe methylation over various regions of the genome and compare the results with other plants. I investigate the density of Non-Differentially Methylated Positions (N-DMPs) across the genome and associate these with certain genes. N-DMPs are positions that are consistently un-methylated (0%), or completely methylated (100%) across all samples. These tend to be found in genes that require strict control of gene expression, such as housekeeping genes that are continuously expressed. I investigate levels of methylation in homeolog pairs (retained from the WGD events) and test whether these are significantly different. I also investigate candidate genes for ADB susceptibility which, although identified using a very small sample, could be useful to further studies into markers associated with susceptibility.

7.1.3 Bisulphite sequencing and alignment software

Methylated cytosines cannot be distinguished from unmethylated cytosines using normal DNA preparation methods. Instead, a bisulphite conversion step must be performed before sequencing. This process converts unmethylated cytosines into uracil residues by deamination (removal of an amino group), while leaving methylated cytosines intact. Upon PCR amplification of DNA, uracil is converted to thymine. Therefore the overall result from the bisulphite conversion is that unmethylated cytosine is turned into thymine, shown in Fig. 7.1. However, the conversion procedure is not perfect; usually around 1% of unmethylated cytosines do not get converted. Therefore the reaction can be performed twice, or steps can be taken later on to correct for the level of conversion efficiency. Sequencing of bisulphite-converted DNA is exactly the same as for normal DNA. However, quality is often reduced by the high salt concentration of the bisulphite reaction.

As many cytosine bases are replaced with thymine, the resulting sequence will have diverged from that of the reference. Alignment of these reads using normal read aligners will be inaccurate as any converted bases will be treated as mismatches and therefore, alignment

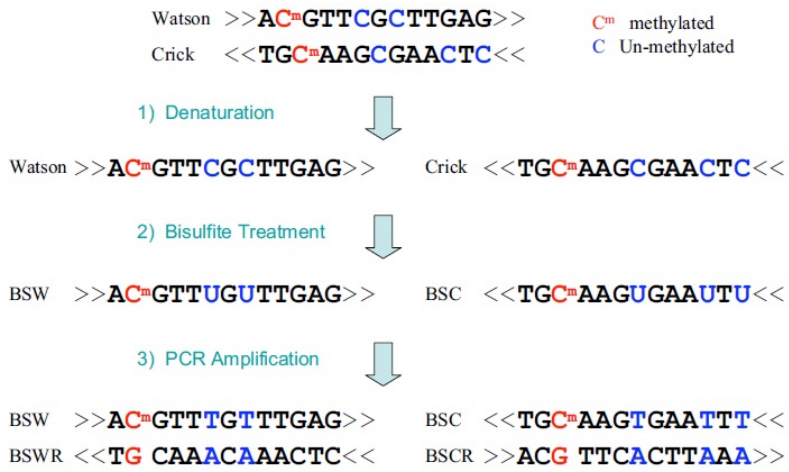


Figure 7.1: Process of bisulphite conversion of unmethylated cytosine into uracil, and then to thymine after PCR amplification. Image taken from Xi & Li (2009)

scores will be vastly reduced. Instead, a bisulphite read aligner must be used, which takes C-T mismatches into account, as well as G-A mismatches on the complementary strand. There are currently two alignment methods used in the vast majority of bisulphite aligners: three-letter alignment and wild-card alignment. These have different ways of handling the mismatches between the reads and reference genome caused by the bisulphite treatment. Wild-card aligners convert all Cs in the reference genome sequence to the wild card letter 'Y', to which both C and T can map. In contrast, three-letter alignment converts all Cs to Ts in the reads (Cs left in the reads must be those that are methylated, as they escaped bisulphite conversion) as well as in the reference genome. Only 3 DNA letters are therefore present (A, G and T), and the converted reads will match perfectly to the reference (the only mismatches remaining would be due to genetic SNPs). The converted Cs in both the reads and the reference are then retrieved after alignment, and can be used to calculate methylation levels at true cytosine positions. Some inaccuracies can occur with either of these methods [Bock 2012]; wild-card alignment is at risk of over-inflating methylation levels as reads with Ts are more likely to be non-uniquely mapped and are therefore excluded from the alignment. Similarly, three-letter alignment reduces the sequence complexity but also decreases the chance that a read will map uniquely in the genome, therefore genome coverage may be reduced if the aligner does not take into account non-unique reads. Non-unique mapping can be allowed in read alignment, where reads will be mapped to both locations, however this also introduces inaccuracies. These issues can be largely cancelled out by using a long read length, which ensures that the chance of non-unique mapping is very small. Then the biggest consideration in choosing a bisulphite read aligner becomes run-time, memory use and user-friendliness of the software.

The most popular tool to use in bisulphite studies seems to be Bismark [Krueger et al. 2011]; which is a very user-friendly pipeline that starts from read mapping using the three-letter alignment method, and ends with methylation calls for every cytosine position in the genome. Bismark allows paired reads to be mapped *only* when both members of a pair are mapped properly together, and this parameter cannot be disabled. If a reference sequence

is in quite a draft form, with lots of broken scaffolds and large gaps of 'N' bases, it is quite likely that many pairs will not map to their correct distances, or that one member may reside in a gap. This could potentially waste many good reads that would map perfectly fine if they were treated as singletons. The CLC bisulphite mapper also uses three-letter alignment but allows for paired reads to be broken. This tool was still in development in 2014 and the rest of the pipeline had not been produced yet. Therefore, I did not use this tool in my study. BSMAP [Xi & Li 2009] was an older program that is still being continuously developed. Therefore, it is up to date with current technology. Although its default method is to use wild-card alignment, it has a three-letter alignment option which can be enabled to make the results as comparable to many other tools as possible. However, BSMAP still allows for paired reads to be broken. The software is both open-source and published [Xi & Li 2009], and downstream tools have already been developed such as a methylation caller, which means that the user does not have to string together several different tools. In addition, BSMAP takes steps to correct for methylated loci that coincide with a C>T SNP, which could cause inaccurate methylation calculations if not corrected for. These can be distinguished using information on the complementary strand; at a completely unmethylated locus, all the Cs should be converted to Ts by the bisulphite reaction, but the complementary base will remain as 100% G. On the other hand, if a heterozygous C>T SNP is present in the individual, the complementary strand should contain roughly a 50:50 ratio of A and G nucleotides. The frequency of A nucleotides on the complementary strand is taken into account when BSMAP calculates the methylation level at every cytosine position, and the methylation level adjusted accordingly.

7.2 Methods

7.2.1 Description of samples and genotypes

In this study, I used a set of twenty ash samples of varying species, genotypes and ADB susceptibility. All of these samples were generated by Erik Kjær's team at the University of Copenhagen, and were grown in a common greenhouse environment. Three samples form clones of one *F. mandshurica* genotype, an Asian ash species with natural low susceptibility to ADB. The remaining 17 samples are from four *F. excelsior* genotypes, which have been described previously in McKinney et al. (2011); two samples of clone number 27, and five each of clone numbers 33, 35 and 40. Samples 33 and 35 were shown to have the least damage from ADB (i.e. low susceptibility) in trials from 2007-2009, and samples 27 and 40 were among those with the most damage (high susceptibility). A description of the samples' origins and phenotypes are in Table 7.1. All samples derive from trees that have been naturally exposed to *H. fraxineus*, however the samples used were not infected themselves.

Sample generation was carried out by members of Erik Kjær's team at the University of Copenhagen (methods are described in Mckinney et al. (2011) and Mckinney et al. (2012)). The samples originate from a population of 40 mature trees, selected from Danish forests between 1934 and 1997. Approximately 25 ramets per tree were grafted onto rootstock in 1998 at two sites in Denmark for field trials; this was prior to ADB arriving in Denmark.

Table 7.1: Description of twenty ash trees with five genotypes used in Bisulphite study. Clones 33 and 35 are considered the low susceptibility *F. excelsior* trees due to low ADB damage in 2009 [McKinney et al. 2011]. Original source location of *F. mandshurica* is unknown, but is presumably in Asia, and ADB damage would be minimal due to natural resistance.

Genotype	Samples	Source location, lat, long	ADB damage in 2009
F. mandshurica	F.mand-1, F.mand-2, F.mand-3		
Clone 27	F.exc 27-1, F.exc 27-2	Helved, 55.0094, 9.9391	>90% samples with >50% damage
Clone 33	F.exc 33-1, F.exc 33-2, F.exc 33-3, F.exc 33-4, F.exc 33-5	Boller, 55.8343, 9.9178	70% samples with <10% damage
Clone 35	F.exc 35-1, F.exc 35-2, F.exc 35-3, F.exc 35-4, F.exc 35-5	Sorø, 55.3855, 11.5851	90% samples with <10% damage
Clone 40	F.exc 40-1, F.exc 40-2, F.exc 40-3, F.exc 40-4, F.exc 40-5	Sorø, 55.527565, 11.736864	>95% samples with >50% damage

The first assessment of ADB damage was performed in 2007, and this was repeated in 2008 and 2009. Additional ramets from eight genotypes (four with the highest and four with the lowest damage according to the assessment in 2009) were selected for inoculation with *Hymenoscyphus fraxineus*. Inoculation was performed in September 2009 using plugs infected with cultures of the pathogen.

7.2.2 DNA extraction, bisulphite conversion and sequencing

DNA was extracted from the twenty samples in December 2013 at the University of Copenhagen, using Qiagen DNeasy Plant Mini Kit (Qiagen, Hilden, Germany). DNA was sent to the Genome Centre at QMUL where the DNA was purified and bisulphite-converted using the Zymo EZ DNA Methylation Gold kit. Libraries were prepared using Epicentre's EpiGnome Methyl-seq kit (Illumina, San Diego, USA), and sequencing was carried out on an Illumina HiSeq, using 2 x 101 bp paired reads.

7.2.3 Data QC and read mapping

Before analyses could be undertaken, quality control steps were taken to process the raw data. Trimming and filtering were performed using the CLC Genomics Workbench v7.5. I tested various read mapping software to perform alignment to the reference genome. Bismark v0.15 [Krueger et al. 2011] resulted in a very low percentage of reads mapping (40-60%) due to excluding broken paired reads. The CLC bisulphite tool mapped between 80 and 90% of reads, depending on parameters, confirming that these broken paired reads did not need to be thrown away. However as previously mentioned, the CLC bisulphite pipeline was still under development in 2014. Therefore, I performed the final mapping to the reference

genome using BSMAP v2.90 [Xi & Li 2009] with the following options changed from the default: -3 (uses a three-letter method for alignment) -w 20 (considers only 20 equal best hits for each read) -g 3 (allows gaps of up to 3 nucleotides). These alignments were used for all further analyses. Using the methratio.py python script included in the BSMAP package, duplicate reads were removed and methylation levels were calculated for all cytosines with strand coverage of at least four reads, using: C_m / (C_m + C_u), where C_m is the number of reads supporting a methylated cytosine, and C_u is the number of reads supporting an unmethylated cytosine. False positive methylation levels obtained from the unmethylated chloroplast genome were used to calculate the efficiency of bisulphite conversion. From all the chloroplast cytosines examined, the conversion efficiency (%) was calculated as 100*(1-(methylated reads / total reads)). All positions that did not have zero methylated coverage were tested for significance using a binomial test to identify positions with 'false positive' methylation levels caused by the non-conversion of unmethylated cytosine during the bisulphite conversion stage. A binomial test compares observed values to an expected success rate, which in this binomial test is the frequency of non-conversion, i.e. (1- (conversion efficiency/100)). These positions should therefore be considered as completely unmethylated, instead of contributing their methylation level, albeit being very low, to any further calculations [Schultz et al. 2012]. The binomial test was performed using a Perl script with the 'pbinom' function from the Math package, with false positive rate as specified above. P-values then corrected for multiple tests using the Benjamini-Hochberg method (Lister et al. 2008) of the 'p.adjust' function in R v 2.15.2. Where cytosines had sufficient methylated coverage but had FDR >0.05 (i.e. not significantly different from the null hypothesis of unmethylated), the number of methylated reads supporting these positions was then set to zero (in only the one sample being analyzed), so as not to support any methylated reads to further calculations [Schultz et al. 2012]. Cytosines at known C->T or G->A SNP loci were filtered out of all files, using 5.1 million polymorphic positions obtained from the range-wide *F. excelsior* diversity panel described in Chapter 6. For some analyses involving calculating averages across regions, weighted methylation levels were used instead of the mean, to adjust the weight that each positions gives to the average calculation based on its coverage [Schulz et al. 2012]. Weighted methylation levels, as described in Schultz et al. (2012), use the following calculation: $\sum C_m$ / $\sum(C_m + C_u)$, so that the contribution of each loci's methylation level to the average value is weighted by its read depth.

All further analyses were performed using custom Perl scripts and results were plotted either in Tableau v9.3, or in R v2.15.2. The genome annotation file "Fraxinus_excelsior_38873_TGAC_v2.gff3" (available at ashgenome.org) was used to define gene, intron, exon and UTR regions, and the file "Fraxinus_excelsior_38873_TGAC_v2.possible_transposable_elements.txt" was used to define which genes were transposable elements.

7.2.4 Data analysis methods

Differential methylation between homeologs derived from two WGD events was investigated to search for evidence of unequal expression. Pairs of homeologs were extracted from the complete list of paralogs identified in Chapter 5 based on their Ks value (0.2-0.4, and 0.5-0.8).

Using logit-transformed mean methylation levels across these homeologs, I investigated differential methylation between the gene pairs using a linear modelling approach (R function 'lm') to test for significant differences. Only genes that contained at least ten sufficiently covered cytosines (those positions covered by >3 reads) were used in this test. The logit transformation was used in order to ensure symmetry of change in the proportions measured between 0 and 1; the output being a numerical value corresponding to a sigmoid curve, where a proportion of 0.5 gives a value of 0, proportions of <0.5 give negative values, and proportions of >0.5 give positive values. Proportions of exactly 0 and 1 were adjusted by 0.001 in order to prevent their logit-transformed values being infinity. All p-values resulting from the linear model tests were then corrected for multiple tests using the Benjamini-Hochberg method of the 'p.adjust' function in R. The linear model test was performed for each *F. excelsior* sample independently. A power calculation was used to investigate power in relation to effect size (i.e. difference in methylation level) and to number of observations (i.e. number of cytosines per gene). The power test was performed using pwr.anova.test function in the 'pwr' package in R v3.3.3, with k (number of groups) = 2, sig.level=0.05, n (number of observations) varying between 5 and 100, and f (effect size) varying between 0 and 1.2.

To investigate how different the methylation patterns of *F. mandshurica* and *F. excelsior* are, I used three different methods to cluster all the samples into groups; Principal Components Analysis (PCA), hierarchical clustering, and Pearson's correlation coefficient. All three methods used a core set of 400,000 cytosines positions that were covered by at least 10 reads in all samples. By using a smaller set of positions, run time and memory usage is reduced whilst still retaining the overall picture of methylation differences and also excluding missing data where some samples lack read coverage. These positions were extracted from all samples using a custom Perl script and formatted to fit the three R functions: 'prcomp', 'dist', and 'cor'. Methylation values were once again adjusted using the logit transformation ('logit' function in R). For PCA, 'prcomp' was used with default parameters. To obtain a distance matrix, 'dist' function was used with 'method = "Euclidean"', and the hierarchical clustering was performed using 'hclust' with 'method = "complete"'. A correlation matrix was made using the 'cor' function in R (without a logit transform of the methylation values), with 'method = "pearson"'.

To investigate possible links between DNA methylation and ADB susceptibility I used two methods. Firstly, I investigated methylation patterns in twenty genes that were already found to have expression levels associated with ADB. This work was performed by my collaborator Andrea Harper and has been described in our joint publication [Sollars et al. 2017]. I calculated the weighted methylation values across the twenty genes for each sample and then tested for differential methylation between the two groups (low versus high susceptibility) using a t-test for each gene.

Secondly, to search for regions in the genome that are differentially methylated between the high and low susceptibility samples, I used the program metilene v0.2-6 [Juhling et al. 2016]. Metilene detects Differentially Methylated Regions (DMRs) between groups of samples using a segmentation algorithm. It first segments the genome into regions with sufficient methylation information, and then scans these regions for pairs of change points

in methylation levels. This ensures that the selected regions have mostly homogeneous methylation levels throughout. A change point is selected where the mean methylation between the two groups attains a maximal change. Intervals are then tested between the two groups using a 2D-KS test, and p-values reported are Bonferroni-corrected for multiple tests.

Metilene is not specific to any cytosine context (despite the paper and manual focusing on CpG methylation), as the input is a matrix of cytosine positions and their methylation values for each sample, with group membership identified within the sample names. It can also handle missing data (i.e. due to low coverage). Compared to other DMR-detecting software, metilene uses less memory and time, whilst being more sensitive in finding DMRs with lower changes in methylation levels [Juhling et al. 2016]. To run metilene, I generated a matrix of methylation values at every cytosine in the reference sequence for every *F. excelsior* sample with low coverage trees excluded (F.exc 40-1, F.exc 35-5, F.exc 33-5) so as not to skew methylation values for their group. Missing data points (due to low coverage) were filled with a dash ('-') character. Power curves were generated for the KS test used in metilene, with various effect and sample sizes used in this study. Power was calculated using a custom script, based on that described in http://stats.stackexchange.com/questions/143245. In brief, it simulates a KS-test 10,000 times using a user-defined ratio of means between two groups (n1 and n2) of a given size, and calculates power based on the proportion of times the null hypothesis is rejected at p¡0.05. In each case, the mean of n1 was kept the same throughout the tests and the mean of n2 was increased to between 1 and 100 times the mean of n1 to generate different effect sizes.

7.3 Results and Discussion

7.3.1 Landscape of DNA methylation across the ash genome

After bisulphite conversion and sequencing, a total of 2,052 million reads were generated, with an average of 102.6 million reads (12x coverage) per sample (Table 7.2).

Approximately 26.6% of all possible cytosines in the *F. excelsior* genome are in a methylated state. Out of these methylated cytosines, 38.5% are in the CHH context, 35.4% in the CG context, and 26.1% in the CHG context (Fig. 7.2). The higher percentage in the CHH context is due to the higher frequency of cytosines in this sequence context overall. These methylation levels are very similar to model tree species *Populus trichocarpa*, but less so to *Betula platyphylla* (though vascular tissue was used in this case instead of leaf) and fellow asterid tomato. The weighted average methylation level was 76.2% for cytosines in the CG context (i.e. 76.2% of all CG cytosines are methylated), 52.0% in CHG and 13.9% in CHH, which are very similar to those calculated in the tomato methylome [Zhong et al. 2013]. Most cytosines appear to be either highly methylated (90-100%) or have very little methylation (0-10%), with very few cytosines having an intermediate level of methylation (Fig. 7.3).

Sample	Million raw reads	Million reads post-QC	Million mapped reads	Million cytosines covered >3 reads	Conversion Rate (%)	Mean Methylation (%)		
						CG	CHG	CHH
F. mand 1	114.5 (13.0x)	82.2 (8.5x)	61.1 (7.7x)	27.4	98.77	75.16	60.46	16.47
F. mand 2	115.8 (13.2x)	85.8 (8.8x)	65.6 (8.3x)	34.4	98.55	74.60	57.77	13.72
F. mand 3	91.6 (10.4x)	63.4 (6.5x)	46.9 (5.9x)	20.8	98.56	76.58	62.93	15.46
F. mand pooled	**322.0 (36.6x)**	**231.3 (23.8x)**	**173.6 (22.0x)**					
F. exc 27-1	87.3 (9.9x)	61.7 (6.3x)	50.2 (6.4x)	24.9	99.07	82.07	68.03	18.45
F. exc 27-2	104.7 (11.9x)	74.0 (7.6x)	61.4 (7.8x)	29.1	99.12	82.55	68.04	16.87
F. exc 27 pooled	**192.1 (21.8x)**	**135.7 (13.9x)**	**111.6 (14.1x)**					
F. exc 33-1	87.5 (9.9x)	58.5 (6.0x)	49.1 (6.2x)	24.4	99.05	82.63	67.92	17.07
F. exc 33-2	124.9 (14.2x)	85.5 (8.8x)	74.3 (9.4x)	43.3	99.08	81.05	62.64	13.09
F. exc 33-3	111.7 (12.7x)	74.8 (7.7x)	65.0 (8.2x)	35.2	99.27	81.09	64.19	14.51
F. exc 33-4	119.6 (13.6x)	82.4 (8.5x)	68.8 (8.7x)	33.3	98.92	81.77	65.70	17.13
F. exc 33-5	69.8 (7.9x)	46.9 (4.8x)	38.5 (4.9x)	18.4	98.97	83.66	70.55	22.66
F. exc 33 pooled	**513.5 (58.4x)**	**348.0 (35.8x)**	**295.7 (37.4x)**					
F. exc 35-1	116.6 (13.3x)	81.2 (8.4x)	67.1 (8.5x)	31.9	98.98	82.20	66.46	18.06
F. exc 35-2	128.8 (14.6x)	90.4 (9.3x)	74.4 (9.4x)	34.2	99.15	81.55	65.45	17.97
F. exc 35-3	112.0 (12.7x)	81.0 (8.4x)	68.4 (8.7x)	31.9	99.26	80.47	63.27	14.04
F. exc 35-4	106.4 (12.2x)	70.4 (7.3x)	57.4 (7.3x)	23.4	99.29	79.73	64.75	16.17
F. exc 35-5	61.2 (7.0x)	39.4 (4.1x)	31.8 (4.0x)	13.3	99.17	81.02	68.79	19.54
F. exc 35 pooled	**524.9 (59.7x)**	**362.5 (37.3x)**	**299.2 (38.0x)**					
F. exc 40-1	41.2 (4.7x)	26.1 (2.7x)	21.8 (2.8x)	10.5	99.03	83.45	71.28	18.02
F. exc 40-2	120.4 (13.7x)	86.4 (8.9x)	70.7 (8.9x)	33.8	99.05	81.54	65.80	19.21
F. exc 40-3	121.0 (13.8x)	83.1 (8.5x)	64.2 (8.1x)	32.3	98.77	81.58	65.18	17.07
F. exc 40-4	113.6 (12.9x)	79.6 (8.2x)	65.0 (8.2x)	32.1	98.59	82.25	66.39	17.14
F. exc 40-5	103.8 (11.8x)	73.4 (7.5x)	60.1 (7.6x)	28.1	99.14	82.41	67.97	20.31
F. exc 40 pooled	**500.1 (56.8x)**	**348.6 (35.8x)**	**281.8 (35.7x)**					
All F. exc pooled	**1730.6 (196.7x)**	**1194.8 (122.8x)**	**988.3 (125.2x)**	**164.4**	**99.23**	**76.18**	**52.04**	**13.92**

Table 7.2: WGBS yield for twenty ash samples. Approximate 880 Mbp genome coverage values in brackets. Coverage value in mapped reads column shows coverage of 718 Mbp non-N genome. Conversion efficiency for each sample was used in separate binomial tests as described in Section 7.2.3. Mean Methylation (%) describes percentage of cytosines in each sequence context that are methylated.

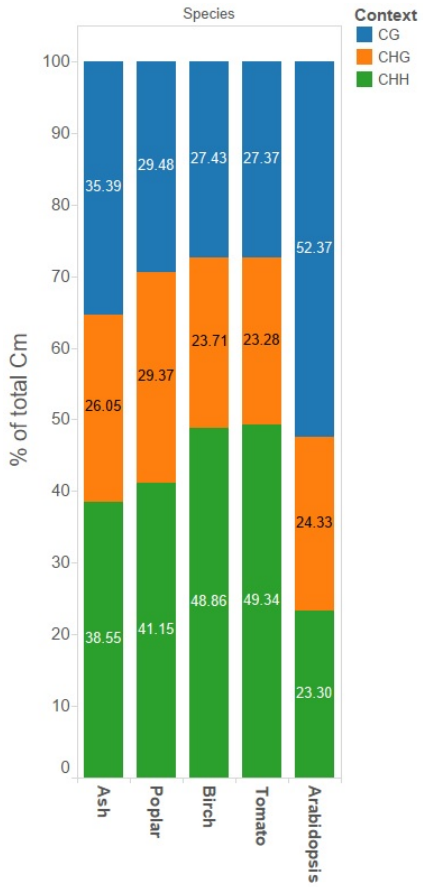


Figure 7.2: Percentage of methylated cytosines in each sequence context; CG, CHG and CHH (where H = A, C, or T), from pooled mapping of all *F. excelsior* trees. Values taken from ash leaf (this study), *Populus trichocarpa* (Poplar) leaf [Liang et al. 2014], *Betula platyphylla* (Birch) vascular tissue [Su et al. 2014], *Solanum lycopersicum* (Tomato) leaf [Zhong et al. 2013] and *Arabidopsis thaliana* leaf [Zhong et al. 2013].

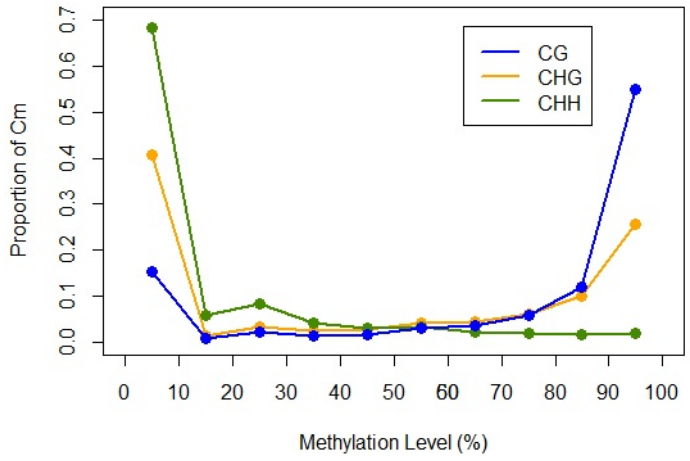


Figure 7.3: Proportion of methylated bases in each context occurring at various methylation levels. Most CHH cytosines are unmethylated or methylated at very low levels, while most CG cytosines are methylated at very high levels. Relatively few cytosines are methylated at medium (20-80%) levels.

The structural annotation file "Fraxinus_excelsior_38873_TGAC_v2.gff3" (available at ashgenome.org), was used to investigate methylation across different genomic regions. Methylation in gene regions was lower than the genomic average in all sequence contexts, but especially so in the CHG context (Fig 7.4). Introns were methylated to a higher degree than exons, and transposable elements (TEs) were substantially more methylated than non-TE genes in all contexts. Sharp dips in methylation level in all contexts can be seen at the start and end sites of genes, especially in the CG context, but not in TE genes where methylation slightly increases (Fig. 7.5).

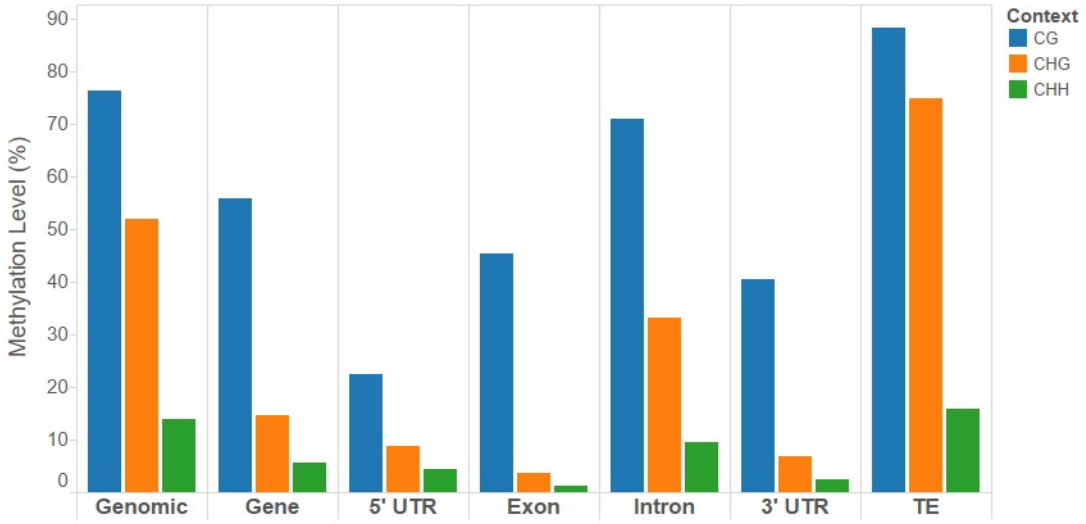


Figure 7.4: Weighted methylation levels across genomic regions in *F. excelsior*.

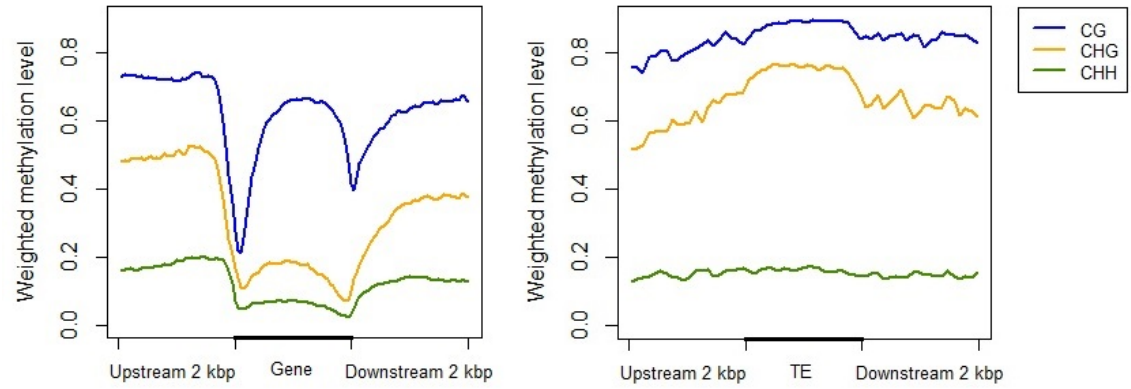


Figure 7.5: Weighted methylation levels across and in 2 kbp flanking regions of genes (left image) and transposable elements (TEs, right image). Each section of genes and flanking regions was split into 40 bins, and of TEs into 20 bins. Methylation across genes dips at the start and end sites, whereas methylation slightly increases across TEs.

Out of 3.1 million cytosines covered in all *F. excelsior* samples, 459,904 (14.5%) were classed as non-differentially methylated positions (N-DMPs), with 97.5% of these being unmethylated (zero methylation), and 2.5% being completely methylated (methylation level of '1' across all samples). Of these 459,904 N-DMPs, 23,710 are in the CG context, 21,454 in the CHG context, and 414,740 in the CHH context. In addition, 73,724 (16.0%) of these N-DMPs occur in gene regions; 5,456 in CG, 6,474 in CHG and 61,794 in the CHH context. Variability in the density of gene N-DMPs exists. Twenty genes with the highest density of N-DMPs relative to their length are listed in Table 7.3, of which all their N-DMPs are completely unmethylated. The majority of these genes have functions that are very conserved across plants and/or eukaryotes, such as those associated with photosynthesis (e.g., NADH hydrogenase subunits), or various ribosomal proteins. This suggests the importance of DNA methylation in regulating gene expression.

Table 7.3: The twenty genes with the highest density of N-DMPs. All N-DMPs were completely unmethylated in all *F. excelsior* samples. Density calculated as: #N-DMPs/(gene length*2) where multiplication by two takes into account both strands of DNA.

Gene ID	N-DMP Density	Gene Function
376950	0.2321	None annotated
016370	0.2119	pg1 protein, lyase activity
376920	0.1712	NADH dehydrogenase subunit 2
016400	0.1689	NADH dehydrogenase subunit 2
376910	0.1668	NADH dehydrogenase subunit 2
009390	0.1602	ATP-ase, AAA-type
016410	0.1558	NADH dehydrogenase subunit 2
238880	0.1545	Ribosomal protein L5
285710	0.1488	Ribosomal protein L2
078310	0.1468	Photosystem II Cp43 Chlorophyll partial
196460	0.1455	Ribosomal protein S12
016420	0.1454	ATP-ase, AAA-type
016390	0.1432	Ribosomal protein S7
277070	0.1429	NADH dehydrogenase subunit 2
376900	0.1424	ATP-ase, AAA-type
363290	0.1418	Ribosomal protein L23
197880	0.1409	NADH dehydrogenase subunit 4
137570	0.1405	Photosystem I p700 apoprotein a1
190110	0.1403	NADH dehydrogenase subunit 2
221320	0.1396	Photosystem I p700 apoprotein a2

7.3.2 Many homeologs retained after WGD are differentially methylated

By extracting the pairs of genes involved in two WGD events and measuring their methylation, I was able to investigate differential methylation between the two copies. As an example, the methylation values of homeologs in one sample (F.exc 33-2) and adjusted p-values from the linear model are shown in Fig. 7.6.

The number of homeologs identified as differentially methylated depended heavily on the overall coverage of the sample, as coverage thresholds were applied before testing. Table 7.4 shows the results for each tree. On average, 23.4% of the homeolog pairs are differentially methylated in the CG context in each sample, with 57/1066 pairs consistently differentially methylated in at least ten samples. In the CHG context, 28.6% of pairs are differentially methylated, 239/1106 across ten samples or more, and in the CHH context, 24.5% of pairs are differentially methylated, 396/1814 across ten samples or more. These values are much higher than other studied plant species, for example Schmitz et al. (2013a) found 602/9793 (6.1%) homeologs in soybean were differentially methylated, and in *Arabidopsis* an even lower proportion of 4/497 (0.8%), although *Arabidopsis* has a much lower number of homeologs altogether.

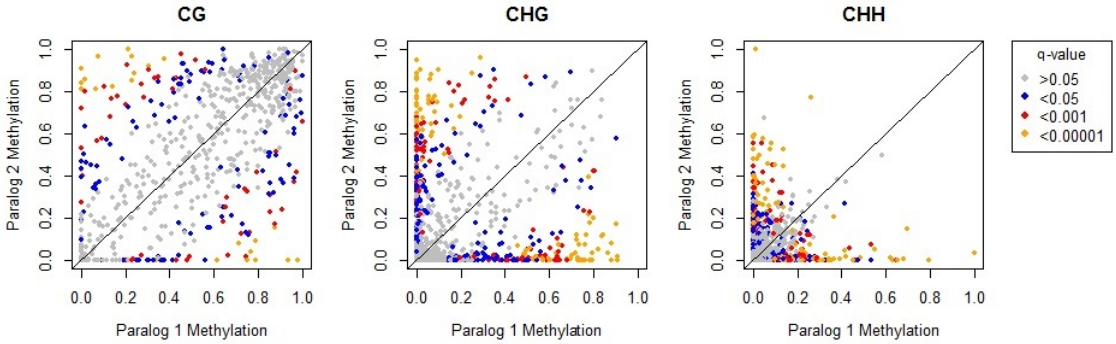


Figure 7.6: Differential methylation in homeologs of sample F.exc 33-2 (as an example), split into CG, CHG and CHH contexts, plotted as methylation of one paralog in a pair against methylation of the other. Q-values shown are FDR-adjusted p-values. Low q-values tend to occur in top-left and bottom-right corners of the graphs (where one paralog has low methylation and the other high methylation) and along each zero axes; where one paralog has zero methylation and the other has at least a medium level of methylation.

Table 7.4: Percentage of homeolog pairs (with at least ten cytosines covered by >3 reads) that are significantly differentially methylated for each *F. excelsior* tree. Samples already identified as low coverage (F.exc 40-1, F.exc 35-5, F.exc 33-5) naturally have lower numbers of homeolog pairs meeting the coverage criteria, therefore percentages for these trees may be skewed by low sample size.

Samples	Significantly differentially methylated gene WGD pairs out of total with sufficient coverage		
	CG	CHG	CHH
F.exc 27-1	128/553 (23.1%)	314/1045 (30.0%)	516/2281 (22.6%)
F.exc 27-2	186/700 (26.6%)	377/1263 (29.8%)	624/2422 (25.8%)
F.exc 33-1	131/475 (27.6%)	251/893 (28.1%)	472/2163 (21.8%)
F.exc 33-2	465/1690 (27.5%)	622/2214 (28.1%)	808/2992 (27.0%)
F.exc 33-3	304/1189 (25.6%)	485/1823 (26.6%)	768/2836 (27.1%)
F.exc 33-4	300/1094 (27.4%)	437/1688 (25.9%)	694/2726 (25.5%)
F.exc 33-5	50/177 (28.2%)	139/410 (33.9%)	371/1577 (23.5%)
F.exc 35-1	212/846 (25.1%)	403/1472 (27.4%)	655/2599 (25.2%)
F.exc 35-2	296/1095 (27.0%)	462/1689 (27.4%)	738/2722 (27.1%)
F.exc 35-3	256/1006 (25.4%)	430/1650 (26.1%)	654/2687 (24.3%)
F.exc 35-4	110/488 (22.5%)	259/933 (27.8%)	536/2240 (23.9%)
F.exc 35-5	15/75 (20.0%)	71/175 (40.6%)	210/930 (22.6%)
F.exc 40-1	7/37 (18.9%)	28/89 (31.5%)	133/672 (19.8%)
F.exc 40-2	279/1070 (26.1%)	486/1709 (18.4%)	719/2745 (26.2%)
F.exc 40-3	254/1032 (24.6%)	447/1671 (26.8%)	629/2747 (22.9%)
F.exc 40-4	250/1028 (24.3%)	444/1622 (27.4%)	647/2699 (24.0%)
F.exc 40-5	175/666 (26.3%)	367/1203 (30.5%)	648/2440 (26.6%)

One difference between the findings in soybean and ash is that in the CG context, methylation of homeologs in soybean were mostly equal, leading the authors to not consider CG methylation in their analysis of differential methylation. CG methylation has also been found to not be as repressive as CHG and CHH methylation, adding another reason for ex-

cluding it. In fact, Wang et al. (2015a) studied methylation in cassava paralogs and found that the gene with higher CG-methylation often had higher expression. However, I have found that very similar proportions of homeologs are differentially methylated regardless of sequence context. Due to the lack of RNA-seq data for these particular twenty samples, I unfortunately cannot tie this in with expression.

One hypothesis for the dissimilar proportions of differentially-methylated homeologs is that the WGD events in the respective species occurred at different times, and therefore the homeologs will have been subject to either more or less restriction by DNA methylation. The most recent WGD event in *Arabidopsis* occurred at around Ks 0.6 [Maere et al. 2005]. Therefore it is slightly older than the recent ash WGD, at around Ks 0.25. The most recent soybean WGD occurred at Ks 0.15 [Schmutz et al. 2010], making it more recent than the ash WGD. Therefore the Ks values do not correlate with the proportion of paralogs that are differentially methylated. It should also be noted that the methods for identifying differential methylated are quite different in the soybean paper. The authors selected any homeolog pair where one had <0.5% methylation and the other >2.5%, while I used a linear modelling approach.

Power curves for this analysis are shown in Fig. 7.7. These indicate the power of rejecting the null hypothesis at various effect sizes between the methylation levels of the two paralogs and with various observations (number of cytosines in this case) per gene. As expected, the power curves show that genes with a higher number of cytosines have more power to detect small effect sizes than genes with fewer cytosines. Genes with fewer cytosines require a larger effect size to reach a high power. In this test, I used genes that had at least 10 cytosines in their sequence, for which power is still relatively high at medium effect sizes (for example, power is around 0.7 for an effect size of 0.6). However most genes will have many more cytosines than ten, therefore the power for detecting small effect sizes in these genes will be even higher.

Homeolog pairs with the lowest p-values from the test for differential methylation across multiple *F. excelsior* samples are shown in Table 7.5. All pairs of genes shown had the same direction of methylation difference across all samples, regardless of significant level, except for one: F.exc 33-5 showed higher methylation in 150960 (difference between methylation levels was not significant in this sample), whereas all other samples showed much higher methylation in 150960. The inconsistency with F.exc 33-5 is likely due to low coverage leading to skewed methylation values and a non-significant result for this particular pair of genes. Although there does not seem to be any particular group or family of genes enriched in this set, a couple are worth mentioning. Firstly, that the pair 133990-299030 (kinesin-like proteins) is one of the most significantly differentially methylated pairs for both the CG and CHG context, and is only a little further down the list for CHH (not shown on table, but in most samples within the top 20). Secondly, there are two different pairs that have been annotated as FAD (flavin adenine dinucleotide)-linked oxidases, one appearing in the CG list and the other in the CHG list. An FAD-linked oxidase has a covalently linked FAD group within the protein, which is used as a co-actor by many enzymes particularly in metabolic processes (e.g., electron transport chain of respiration). It could perhaps be important to

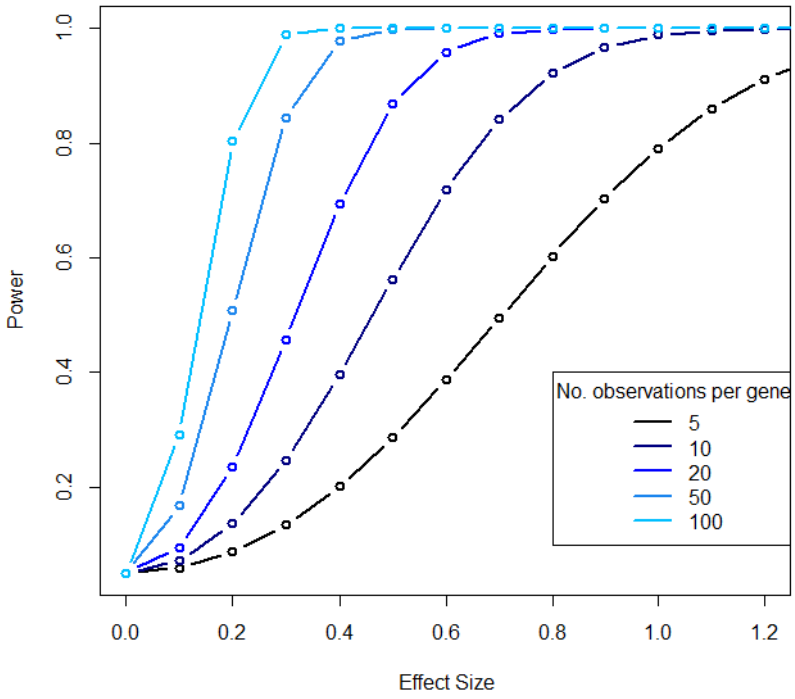


Figure 7.7: Power curves for detecting methylation differences between heomeologs. Generated using the pwr.anova.test function in the 'pwr' R package, for between 5 and 100 cytosines per gene (one gene tested at a time), and for effect sizes between 0 and 1.2, where effect size is calculated as the difference in sample means / standard deviation.

regulate the expression of this protein in order to regulate the metabolic processes it is involved in. Notably, another homeolog pair in the CG list is a component of pyruvate dehydrogenase, a protein also involved in metabolic pathways. This suggests that pathways which tend to be highly conserved across both plants and animals also employ strategies such as DNA methylation to regulate the expression of some of their associated proteins.

Table 7.5: Most consistently differentially methylated homeologs. Counts depict how many times the pair is present in the lowest ten p-values across all *F. excelsior* samples. Note that the table is split into CG, CHG, and CHH contexts (CHG and CHH continued onto next page). All pairs of genes showed the same direction of methylation difference in all samples with significant methylation difference.

Gene 1	Gene 2	Count	Function	GO terms
CG				
150950	150960	13	FAD-linked oxidase, N-terminal	Several, e.g. GO:0050660 (MF: Flavin-adenine dinucleotide binding), GO:0006950 (BP: response to stress), GO:0006040 (BP: amino sugar metabolic process)
133990	299030	10	kinesin-like protein, atp-binding	Several, e.g. GO:0007018 (BP: microtubule based movement), GO:0008569 (MF: ATP-dependent microtubule motor activity, GO:0005871 (CC: kinesin complex)
146520	329760	9	dihydrolipoyllysine-residue acetyltransferase component of pyruvate dehydrogenase mitochondrial-like	Several, e.g. GO:0009941 (CC: chloroplast envelope), GO:0022626 (CC: cytosolic ribosome), GO:0016746 (MF: transferase activity)
227480	280260	6	mfp1 attachment factor 1-like	GO:0005515 (MF: protein binding), GO:0043231 (CC: intracellurar membrane-bound organelle)
285590	398790	5	ring-h2 finger protein	GO:0008270 (MF: zinc ion binding)
169730	353730	5	cationic amino acid transporter	Several, e.g. GO:0003333 (BP: amino acid transmembrane transport), GO:0005886 (CC: plasma membrane)
115990	166400	5	inter-alpha-trypsin inhibitor heavy	GO:0016020 (CC: membrane)
113170	126410	5	anthocyanin 5-aromatic	GO:0047672 (MF: anthranilate N-benzoyltransferase activity), GO:0080167 (BP: response to karrikin (plant growth regulator))
159580	240960	4	laccase 110am multicopper oxidase	Several e.g. GO:0046688 (BP: response to copper ion), GO:0005507 (MF: copper ion binding), GO:0052716 (MF: hydroquinone:oxygen oxidoreductase activity),
250570	373520	4	snrk1-interacting protein 1	GO:0009507 (CC:chloroplast)
CHG				
094250	368880	17	atpase splayed	GO:0003677(MF: DNA binding), GO:0005524(MF: ATP binding), GO:0004386 (MF: helicase activity)
133990	299030	17	kinesin-like protein, atp-binding	Several, e.g. GO:0007018 (BP: microtubule based movement), GO:0008569 (MF: ATP-dependent microtubule motor activity, GO:0005871 (CC: kinesin complex)
026950	114060	15	None annotated	GO:0003676(MF: nucleic acid binding), GO:0003723 (MF: RNA binding), GO:0006396 (BP: RNA processing)
307070	375990	14	Zinc finger, PHD-type	GO:0008270 (MF: zinc ion binding)
013250	272830	13	None annotated	No GO terms. IPR 020864 (Membrane attack complex component/perforin (MACPF) domain)
339290	380270	12	FAD-linked oxidase	Several, e.g. GO:0050660(MF: flavin adenine dinucleotide binding), GO:0006979(BP: response to oxidative stress), GO:0009793 (BP: embryo development), GO:0010197 (BP: polar nucleus fusion)
327050	381680	12	cytosolic fe-s cluster assembly factor nbp35	GO:0005524 (MF: ATP binding), GO:0042803 (MF: protein homodimerization activity), GO:0051536(MF: iron-sulfur cluster binding)
155290	378020	11	cytosolic delta subunit	Several e.g. GO:0006457 (BP: protein folding), GO:0046686 (BP: response to cadmium ion), GO:0005524 (MF: ATP binding)
024730	073330	9	regulatory-associated protein of tor 1-like	GO:0009793 (BP: embryo development), GO:0016049 (BP: cell growth)
383510	395700	8	None annotated	GO:0005515 (MF: protein binding), IPR017986 (WD40-repeat-containing domain), IPR006594 (LisH homology motif)

CHH results on next page

Gene 1	Gene 2	Count	Function	GO terms
CHH				
030090	095610	15	None annotated	GO:0003755 (MF: peptidyl-prolyl cis-trans isomerase activity), GO:0006457 (BP: protein folding)
055530	207470	13	mediator of rna polymerase ii transcription subunit	GO:0001104 (MF: RNA polymerase II transcription cofactor activity), GO:0016592 (CC: mediator complex)
091230	270880	10	actin associated protein	GO:0009536 (CC: plastid)
208290	254280	10	nucleoprotein tpr	GO:0044699 (BP: single organism process), GO:0044428 (CC: nuclear part), GO:0009987 (BP: cellular process)
105380	292350	10	p-loop containing ntpase domain-containing protein	GO:0005524 (MF: ATP binding), GO:0017111 (MF: nucleoside-triphosphatase activity), GO:0005739 (CC: mitochondrion)
163800	175560	8	udp-n-acetylmuramoyl-l-alanyl-d-glutamate– - diaminopimelate ligase-like	Several, e.g. GO:0007049 (BP: cell cycle), GO:0009658 (BP: chloroplast organisation), GO:0009252 (BP: peptidoglycan biosynthetic process)
024720	073340	7	None annotated	GO:0005622 (CC: intracellular), GO:0006886 (BP: intracellular protein transport)
346190	353990	6	None annotated (LRR-containing)	GO:0016301 (MF: kinase activity), GO:0016491 (MF: oxidoreductase activity), GO:0016310 (BP: phosphorylation)
091760	125860	6	None annotated	GO:0055114 (BP: oxidation-reduction process), GO:0010087 (BP: phloem or xylem histogenesis), GO:0032875 (BP: regulation of DNA endoreduplication),

7.3.3 Methylation differences between two *Fraxinus* species and within genotypes

Using Principal Components Analysis (PCA), I was able to distinguish the two species in my samples based on their methylation patterns (Fig. 7.8). The plot of PC1 vs PC2 clearly shows separation of the *F. mandshurica* trees (red crosses) away from all other *F. excelsior* individuals along the PC1 axis, but the *F. excelsior* trees are not separated very much. Using the loadings of each cytosine position for PC1, I was able to obtain a list of the top 40 positions responsible for most of the separation between the two species (Table 7.6). The vast majority of these positions were not within gene regions, with the exception of two: Contig2376 position 21815 is within gene 119700, a phospholipid-translocating ATPase, and Contig8085 position 87294 lies within gene 356660, an RNA recognition motif-containing protein.

There are also three outliers on PC plots 7.8, which are all very low coverage *F. excelsior* trees. The pink triangle outlier on PC3 is F.exc 35-5, the black diamond outlier on PC2 is F.exc 40-1, and the orange diamond outlier on PC4 is F.exc 33-5. All of these samples had an average genome coverage after mapping of <5x. When these three low coverage samples were excluded, the four *F. excelsior* genotypes become distinguishable from each other along the main PCs (fig. 7.9). Firstly, this is evidence for genotype-specific methylation patterns, as samples from the same genotype cluster closer together. There still remains some variation between the clones within each genotype, however. As all trees were grown in a common environment, it is unlikely that environmental effects caused this. Therefore, it is more likely to be epigenetic stochasticity in the trees. This analysis also reinforces the need for sufficient genome coverage, ideally >10x raw read coverage and >5x after QC and mapping, before attempting to compare different methylomes.

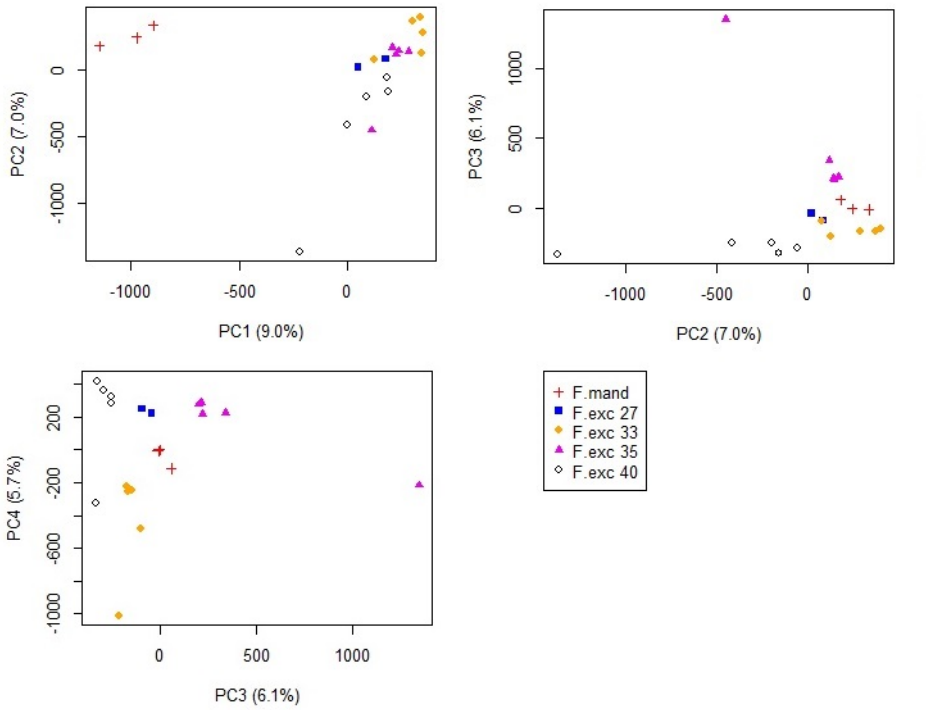


Figure 7.8: Principal Components Analysis of methylation values from 400,000 cytosines across all samples. *F. mandshurica* trees are split from *F. excelsior* trees along PC1. Three outliers represent low coverage samples: pink triangle outlier on PC3 is F.exc 35-5, black diamond outlier on PC2 is F.exc 40-1, and orange diamond outlier on PC4 is F.exc 33-5.

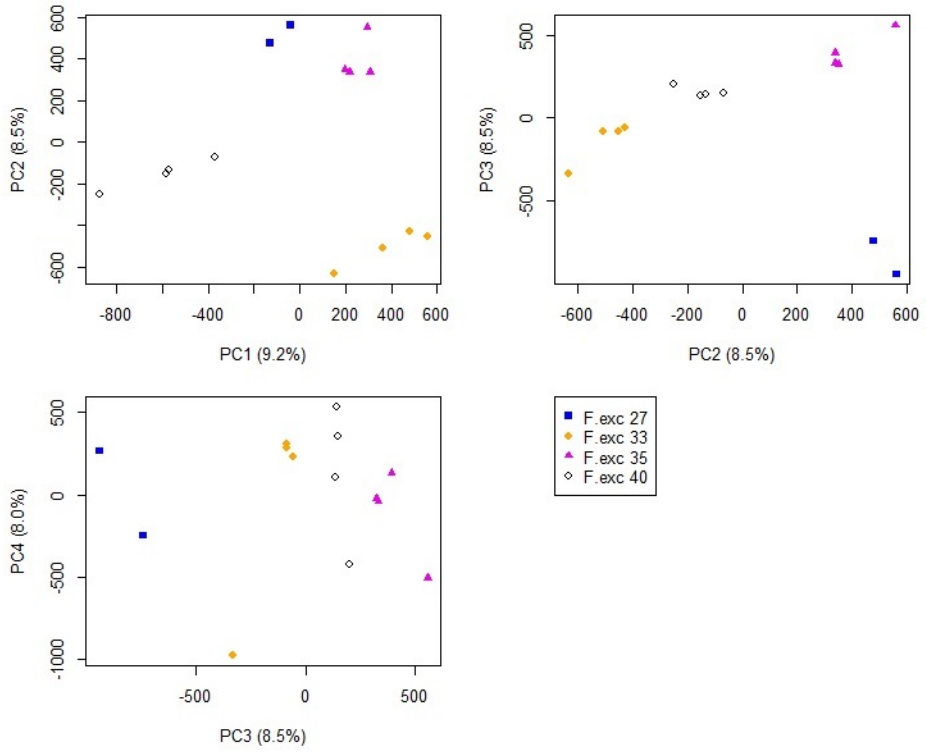


Figure 7.9: Principal Components Analysis of methylation values from 400,000 cytosines for high coverage *F. excelsior* samples. F.exc 33-5, F.exc 35-5, and F.exc 40-1 were removed from the PCA, leaving samples that cluster highly within each genotype. However some variation in methylation is visible between samples within each genotype.

Table 7.6: Genomic positions with most effect on separation of the two *Fraxinus* species, based on loading along PC1. Only two positions were within gene regions: Contig2376 position 21815 is within gene 119700, a phospholipid-translocating ATPase, and Contig 8085 position 87294 lies within gene 356660, an RNA recognition motif-containing protein.

Contig	Position	Context	PC1 Loading	Mean methylation *F. mandshurica*	Mean methylation *F. excelsior*
5381	50342	CHH	-0.0127	1.000	0.046
1525	59199	CG	-0.0120	1.000	0.026
3593	6587	CHH	-0.0118	1.000	0.147
2376	21815	CHH	-0.0116	1.000	0.124
2226	58007	CG	-0.0112	1.000	0.037
1525	59192	CG	-0.0107	0.976	0.022
3593	6605	CHH	-0.0102	0.961	0.062
1742	78238	CHH	-0.0100	0.961	0.047
1525	59187	CG	-0.0099	0.970	0.031
5567	17522	CHH	-0.0098	0.910	0.057
345	160178	CHH	-0.0095	1.000	0.290
1019	198545	CHH	-0.0095	1.000	0.162
537	163284	CHH	-0.0093	0.972	0.081
1183	127103	CHH	-0.0089	0.523	0.000
1525	59185	CHG	-0.0086	0.944	0.023
3339	39370	CHH	-0.0085	0.774	0.190
4458	28642	CHH	-0.0084	0.896	0.035
3147	44856	CHH	-0.0083	0.751	0.043
2376	112909	CHH	-0.0081	1.000	0.222
2271	6442	CHH	-0.0081	0.984	0.076
3225	5301	CG	0.0133	0.000	0.985
2238	83939	CHG	0.0133	0.000	0.963
2503	15395	CHG	0.0129	0.000	0.973
3315	36341	CG	0.0127	0.000	0.987
8085	87294	CG	0.0125	0.000	0.974
1546	63329	CHG	0.0124	0.000	0.969
1226	46428	CG	0.0124	0.000	0.983
83324	24258	CG	0.0123	0.000	0.968
87	308463	CHG	0.0122	0.000	0.967
817	74700	CG	0.0116	0.000	0.939
1013	39297	CHG	0.0114	0.000	0.893
3129	72215	CHG	0.0114	0.000	0.930
958	80855	CHG	0.0109	0.000	0.768
5552	6983	CG	0.0108	0.000	0.930
4818	47459	CG	0.0107	0.091	0.967
1417	7761	CHH	0.0105	0.000	0.719
2271	886	CG	0.0105	0.000	0.956
48564	569	CHG	0.0104	0.072	0.983
16	144183	CG	0.0104	0.000	0.936
3829	27916	CG	0.0104	0.321	0.983

Hierarchical clustering better shows which genotypes' methylation patterns are more similar to each other (Fig. 7.10). As expected the three *F. mandshurica* trees form an outgroup by themselves away from the other *F. excelsior* samples. Without the three low coverage samples described previously, the *F. excelsior* samples group clearly into their genotypes, giving further evidence for genotype-specific methylation patterns, despite a common environment. It is also interesting that the two genotypes with low susceptibility to ADB (33

and 35) cluster together the closest. However, I do not expect that any 'susceptibility loci' lie within the 400,000 cytosine positions surveyed, or if they did, that they alone would cause such a close clustering of the two genotypes. It could purely be by chance that the two closest genotypes happen to be those with low susceptibility, and if more genotypes were added the clustering would likely be different.

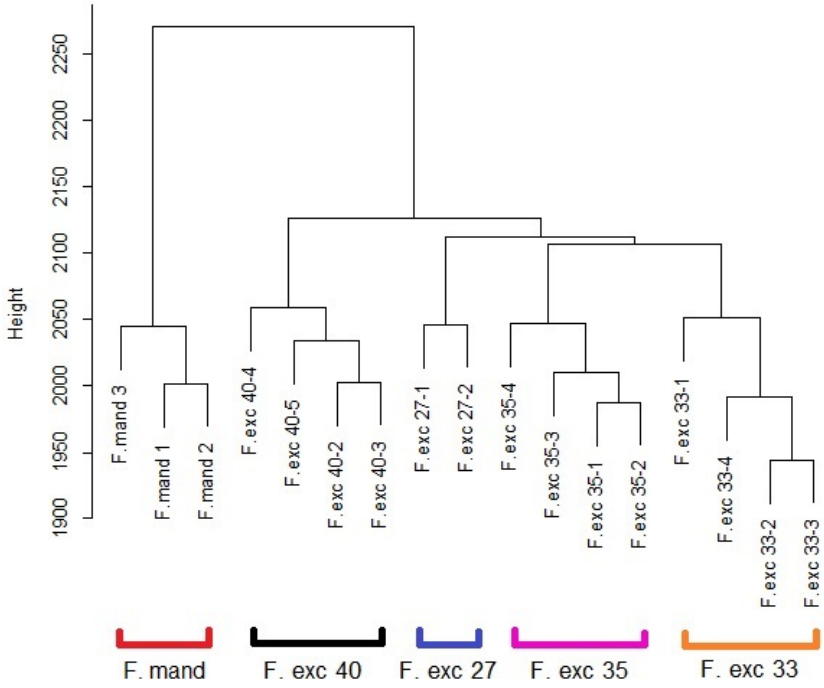


Figure 7.10: Hierarchical clustering of high coverage samples using methylation values from 400000 positions shows that samples cluster according to genotype. All samples shown have average genome coverage >5x. Therefore F.exc 35-5, F.exc 40-1, and F. exc 33-5 were removed. These outlier samples clustered away from all other *F. excelsior* genotypes when included, as if they were outgroups.

To find out the difference in variation between samples within a genotype compared to samples with different genotypes, I performed a Pearson's correlation test of logit-transformed methylation values between every sample. A heatmap of the correlation matrix is shown in Fig. 7.11. The *F. mandshurica* individuals have low correlation values with all other *F. excelsior* trees. The low coverage samples (F.exc 33-5, F.exc 35-5 and F. exc 40-1) also show slightly lower correlation values to other *F. excelsior* genotypes, and their values are in line with their average coverage values (e.g., F.exc 40-1 has the lowest coverage at 2.8x and also has the lowest correlation values to other samples).

Correlation values within genotypes (usually around 0.96) are slightly higher than between genotypes (usually around 0.955), but clearly not by a large amount. Due to being grown in a controlled common environment, the methylomes of different genotypes are likely to be very similar, but with some differences that are genotype-specific. The samples identified as having low coverage (F.exc 33-5, F.exc 35-5, F.exc 40-1) had slightly lower correlation with all other trees (around 0.95), again showing the skewing of methylation values due to

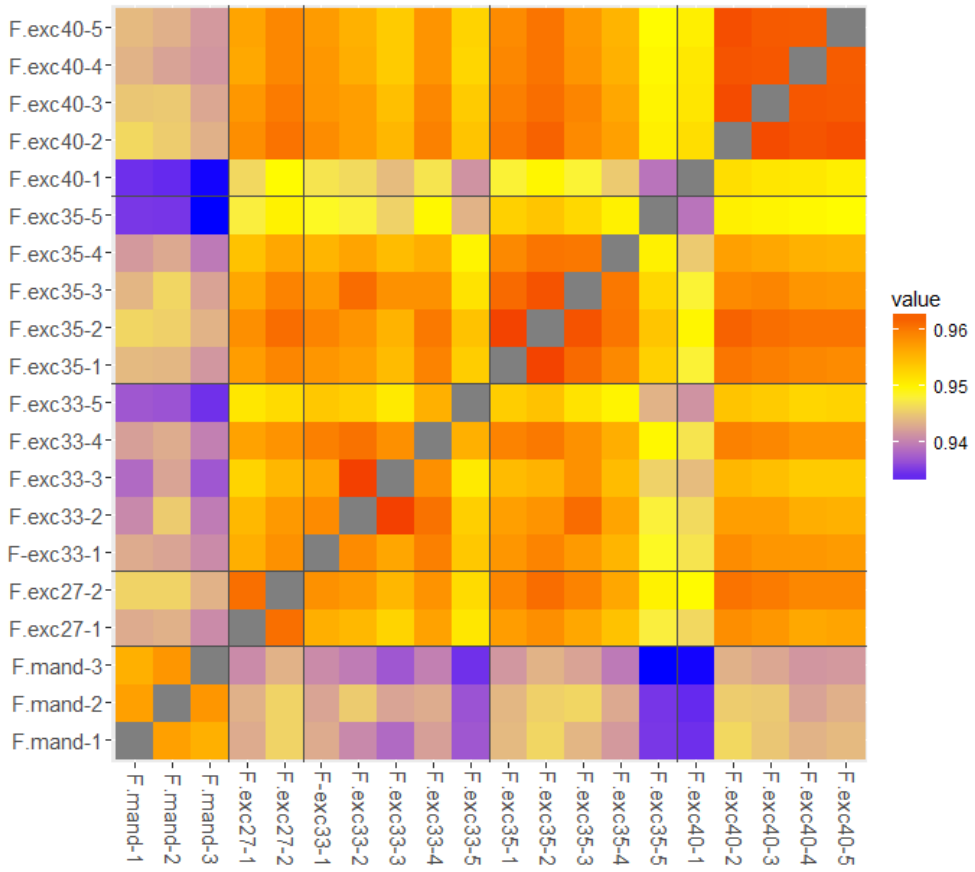


Figure 7.11: Correlation matrix heatmap of all samples using methylation values and Pearson's correlation coefficient values. All samples show very slightly higher correlation within the same genotype (∼0.96), than with samples of a different genotype (∼0.955), reflecting some possible genotype-specific methylation patterns, but there still remains variation between clones within each genotype. *F.mandshurica* samples show lower correlation with all *F. excelsior* (∼0.94). The low coverage *F. excelsior* samples, particularly F.exc 35-5 and F. exc 40-1, also show lower correlation with other *F. excelsior* trees (∼0.95).

low coverage. The difference between the two species is more apparent (around 0.94). This likely reflects species-specific methylation patterns, but some of these differences could also be due to using *F. mandshurica* reads mapped to an *F. excelsior* reference sequence. Of course, differences will be expected between the two genomes (diverged sequences), and this could cause inaccuracies in mapping of reads or a reduction in coverage. However, the genome coverage for all *F. mandshurica* samples was sufficiently high, and by considering only positions where all samples are covered, this removes the possibility of selecting sites where *F. mandshurica* reads have not mapped. In addition, the mapping program BSMAP takes steps towards adjusting for sites that could also be SNPs, by taking the base calls on the complementary strand into account when calculating the methylation value. Therefore, I think that these potential issues have had a negligible effect on the results.

7.3.4 Methylation in genes relating to ADB susceptibility

Despite the low number of samples available, I also made efforts to look for differences in methylation patterns between low and high ADB susceptibility trees. I am aware that twenty samples and four genotypes is limited for an association-type study. However I believe these analyses could be useful as preliminary studies or, at the very least, for a proof-of-concept study.

Firstly, I investigated methylation patterns in twenty genes that were already found to have expression levels associated with ADB. Results are shown in Table 7.7. Two genes were significantly differentially methylated between the two groups; 261470 (soc1-like protein) was significant both in the CG and CHG context, and 178920 (cinnamoyl-CoA reductase 2) was significant in the CHG context (Fig. 7.12). All have higher methylation in the high susceptibility group. However, after adjusting p-values for multiple tests, only 178920 remained significant.

The expression levels of 178920 were previously shown to be higher among high susceptibility samples (different samples to those used here). This does not necessarily fit with the higher CHG methylation levels observed in this study as CHG methylation is usually correlated with suppressed gene expression. The enzyme cinnamoyl-CoA reductase 2 is part of phenylpropanoid biosynthesis. Phenylpropanoids form parts of structural compounds which, among other functions, provide defence against pathogens and herbivores [Konig et al. 2014; Ballester et al. 2013]. Perhaps this enzyme is upregulated in high susceptibility samples, which in turn causes a change in phenylpropanoid synthesis and therefore, in the defence against pathogens such as *Hymenoscyphus fraxineus*.

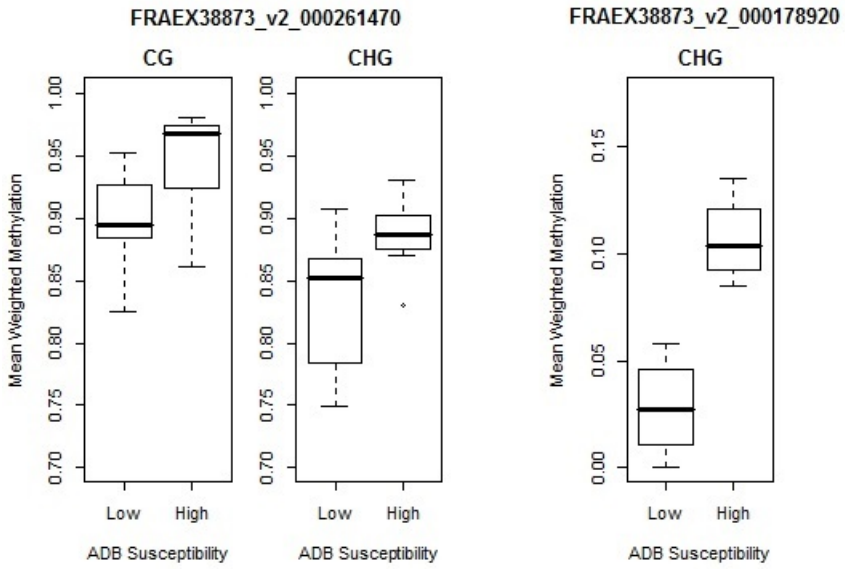


Figure 7.12: Boxplots of mean weighted methylation levels for high and low susceptibility samples, for two genes found to be significantly associated with ADB susceptibility [Sollars et al. *in press*].

Table 7.7: Average weighted methylation levels in twenty genes known to be associated with ADB susceptibility from [Sollars et al. 2017], for all low and high susceptibility samples. 'PD' is short for 'PREDICTED' in second column. Methylation levels were only calculated for genes with at least ten informative cytosines in the sequence context in question, i.e., some genes only have CHH listed as there were <10 CG and CHG informative cytosines. ± denotes standard error.

Gene	Top BLASTX hit	Mean weighted methylation		T-test p-value	FDR-corrected p-value
		Low susceptibility	High susceptibility		
298540	PD: calcium-transporting ATPase 9, plasma membrane-type [Erythranthe guttata]	CG:0.57±0.08 CHG:0.003±0.0018 CHH:0.0047±0.002	CG:0.46±0.06 CHG:0.003±0.0015 CHH:0.007±0.0036	CG:0.308 CHG:0.896 CHH:0.503	CG:0.676 CHG:0.922 CHH:0.744
368850	PD: uncharacterized protein LOC105163554 [Sesamum indicum]	CHG:0.012±0.003 CHH:0.0145±0.005	CHG:0.016±0.009 CHH:0.022±0.014	CHG:0.701 CHH:0.521	CHG:0.841 CHH:0.744
261470	soc1-like protein [Olea europaea]	CG:0.898±0.012 CHG:0.831±0.014 CHH: 0.188±0.012	CG:0.946±0.019 CHG:0.886±0.012 CHH:0.202±0.021	**CG:0.037** **CHG:0.016** CHH:0.537	CG:0.444 CHG:0.288 CHH:0.744
251090	PD: MADS-box protein SVP-like isoform X1 [Sesamum indicum]	CHH:0.026±0.009	CHH:0.047±0.015	CHH:0.223	CHH:0.643
048340	PD: MADS-box protein SVP-like [Sesamum indicum]	CG:0.854±0.03 CHG:0.512±0.046 CHH:0.098±0.009	CG:0.773±0.037 CHG:0.521±0.027 CHH:0.101±0.015	CG:0.300 CHG:0.879 CHH:0.848	CG:0.657 CHG:0.922 CHH:0.922
178910	PD: cinnamoyl-CoA reductase 2 [Sesamum indicum]	CHH:0.122±0.037	CHH:0.078±0.04	CHH:0.442	CHH:0.723
245740	PD: mitochondrial arginine transporter BAC2-like [Solanum pennellii]	CHG:0.129±0.03 CHH:0.155±0.034	CHG:0.095±0.016 CHH:0.157±0.043	CHG:0.405 CHH:0.976	CHG:0.709 CHH:0.976
173540	PD: MADS-box protein SVP-like isoform X1 [Nelumbo nucifera]	CG:0.6410.033 CHG:0.454±0.029 CHH:0.096±0.021	CG:0.697±0.081 CHG:0.412±0.07 CHH:0.057±0.009	CG:0.455 CHG:0.522 CHH:0.106	CG:0.723 CHG:0.723 CHH:0.509
032420	PD: uncharacterized protein LOC105168112 [Sesamum indicum]	CHH:0.101±0.029	CHH:0.192±0.056	CHH:0.131	CHH:0.509
265110	PD: potassium channel KAT3 [Erythranthe guttata]	CHH:0.023±0.007	CHH:0.005±0.003	CHH:0.123	CHH:0.509
262400	PD: serine/threonine-protein kinase AFC2 isoform X1 [Sesamum indicum]	CG:0.595±0.113 CHG:0.015±0.009 CHH:0.006±0.003	CG:0.513±0.059 CHG:0.003±0.003 CHH:0.001±0.0008	CG:0.599 CHG:0.239 CHH:0.112	CG:0.776 CHG:0.643 CHH:0.509
347350	PD: glucuronoxylan 4-O-methyltransferase 1-like [Sesamum indicum]	CHH:0.0025±0.002	CHH:0.003±0.003	CHH:0.858	CHH:0.922
048360	PD: MADS-box protein SVP-like isoform X1 [Sesamum indicum]	CG:0.772±0.038 CHG:0.519±0.052 CHH:0.135±0.024	CG:0.790±0.026 CHG:0.590±0.021 CHH:0.176±0.009	CG:0.737 CHG:0.228 CHH:0.126	CG:0.86 CHG:0.643 CHH:0.509
199890	PD: uncharacterized protein LOC105957157 [Erythranthe guttata]	CHH:0.292±0.053	CHH:0.193±0.092	CHH:0.329	CHH:0.657
178920	PD: cinnamoyl-CoA reductase 2 [Sesamum indicum]	CG:0.574±0.085 CHG:0.028±0.008 CHH:0.018±0.005	CG:0.372±0.028 CHG:0.107±0.011 CHH:0.014±0.006	CG:0.087 **CHG:<0.001** CHH:0.645	CG:0.509 **CHG:0.007** CHH:0.806
249100	PD: MADS-box protein SOC1-like isoform X1 [Sesamum indicum]	CG:0.864±0.018 CHG:0.579±0.051 CHH:0.132±0.013	CG:0.845±0.041 CHG:0.456±0.066 CHH:0.114±0.011	CG:0.634 CHG:0.159 CHH:0.399	CG:0.806 CHG:0.557 CHH:0.709
318210	PD: protein EMBRYONIC FLOWER 1 isoform X1 [Sesamum indicum]	CHG:0.020±0.006 CHH:0.006±0.004	CHG:0.008±0.008 CHH:0.002±0.001	CHG:0.297 CHH:0.338	CHG:0.657 CHH:0.657

To search for regions in the genome that are differentially methylated between the high and low susceptibility samples, I used the program metilene V0.2-6 [Juhling et al. 2016]. In total, 1,683 significant DMRs were found between the low and high susceptibility *F. excelsior* samples at p<0.001, with 103 in the CG, 112 in the CHG and 1,468 in the CHH context. 20.2% of these overlapped with genes. However, there were stark differences in the

level of gene association between the three sequence contexts. 41/103 (39.8%) CG-DMRs overlapped with a gene, compared to 70/112 (62.5%) CHG-DMRs and 169/1168 (14.5%) CHH-DMRs. The genes associated with the most significant DMRs in each sequence context are listed in Table 7.8 (at end of chapter). However, there are many more significant DMRs that do not overlap genes. Out of these DMR-genes, three have both CG- and CHG-DMRs, meaning that they may be consistently differentially methylated between low and high susceptibility trees in both the CG and CHG contexts. These genes are 043910, which unfortunately has no annotation, 137200 (mechanosensitive ion channel domain-containing protein, MscS), and 197890, also with no annotation.

Although the MscS family of proteins would at first seem to have no connection with disease resistance or response, upon further reading, a tentative link can be made. MscS proteins are present in the membrane (Class I in the cell membrane, and Class II in organellar membranes), and are used to "release osmolytes and prevent cell lysis during hypoosmotic stress", or also act as signal transduction mechanisms during osmotic stress [Wilson et al. 2013]. They respond when the cell is in osmotic stress through mechanical tension changes in the membrane. ADB is thought to cause osmotic stress in cells by causing cell lysis in, among other tissues, the xylem vessels. Perhaps, and I am speculating here, the low susceptibility trees (which have an extremely hypomethylated 137200 gene compared to the high susceptibility samples), have increased expression of this MscS protein, and can therefore cope better with the osmotic stress they are put under during ADB infection. This could allow cells to maintain turgor and prevent cell death. I therefore suggest this as a candidate locus for further study. None of the twenty genes that have expression levels associated with ADB susceptibility (described in Table 7.7) had significant DMRs in any sequence context.

The power curves for this DMR analysis are shown in Fig. 7.13, displaying the power for rejecting the null hypothesis at $p<0.05$ for various sample sizes and difference in means between the two groups. The curves show that power is increased with the larger sample sizes when *F. mandshurica* samples are included. Power is increased further still when the larger group mean (low susceptibility group) is increased in relation to the smaller group. However with the lower sample sizes, power is increased with the opposite scenario; when the mean of the small group is increased in relation to the larger group. Nevertheless, these are very small differences in power when changing the sample size and it should still be noted that power is still very low overall due to the small number of samples per group. Only when the difference between the means of the two groups reaches in the order of a 10x difference, does the power reach over 0.8. Therefore, in order to routinely detect smaller differences between low and high susceptibility trees a much larger sample size would be needed.

7.4 Conclusion

In keeping with previous studies of DNA methylation in plants, I find comparable levels of methylation in each sequence context, and that methylation is increased in transposable elements in comparison to normal, non-mobile genes. I find that completely unmethylated cytosines are enriched in housekeeping type genes that require constant levels of gene

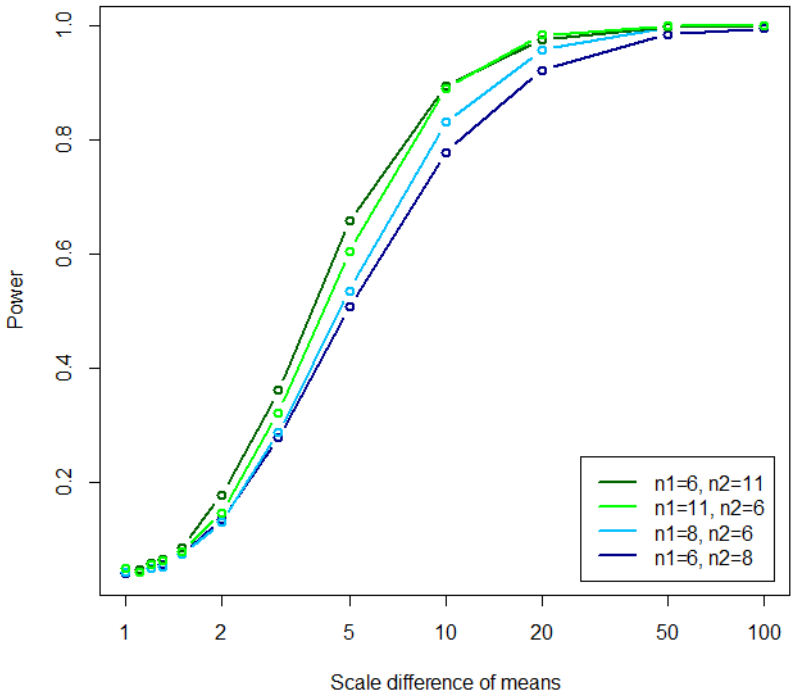


Figure 7.13: Power curves for KS-test used in the metilene program for DMR detection, for various sample effect sizes. Green lines represent tests including *F. mandshurica* samples, with size of low susceptibility group increased from 8 to 11. In each case, mean of n1 is kept the same, while mean of n2 is increased by scale denoted on x-axis.

expression, such as those involved in photosynthesis pathways. Differentially methylated homeologs retained from two recent WGD events are found more frequently in comparison to other studied plants, but this cannot be attributed to the time since the WGD event occurred. Some gene and region candidates for ADB susceptibility are identified from comparisons between the low and high susceptibility groups of trees. However with such a small number of samples these will definitely need to be confirmed in a larger study. Species- and genotype-specific methylation is evident as the the twenty samples cluster first according to species and then, in *F. excelsior* trees, according to genotype. However, methylation is still variable even among genetically identical individuals raised in a common controlled environment.

The implications of these results are that genetically identical individuals still display variability in their epigenome, even when raised in a common environment. This could be stochasticity that may not have an effect on growth or survival of the plant, or it could provide a kind of phenotypic plasticity to the population, creating diversity even when the genetic background is lacking in variability. This could provide scope for adaptation to biotic or abiotic change outside of controlled environments.

Future studies on DNA methylation in ash could follow up on the candidate genes identified as possibly linked to ADB susceptibility. A larger sample size would be needed, so that the genetic background is diverse and the effect of stochasticity reduced. Targeted meth-

ods of Bisulphite-seq would help to reduce the cost of a large-scale study, such as Reduced Representation Bisulphite Seq (RRBS). These techniques would also ensure high coverage at the specific loci, as I have shown that low read coverage skews the methylation values calculated and therefore the interpretation of results. Adding RNA-seq data would also provide additional assurance that any candidate DNA methylation loci or genes identified do actually have an effect on gene expression.

In addition, a more in-depth functional annotation of the homeologs that are differentially methylated could be undertaken. The gene balance hypothesis suggests that certain types of genes may be preferentially retained after duplication events. In general, proteins that form parts of protein complexes or pathways are required to have similar expression levels to the other members of the complex / pathway, and housekeeping genes also need to maintain a steady expression level [Birchler & Veitia 2007; Freeling 2009; Schnable et al. 2011a]. In terms of DNA methylation, these dosage-sensitive genes may be more likely to have similar methylation levels so that balanced expression is maintained and the gene dosage in complex pathways does not become disrupted. It would be interesting to use the annotations of the homeologs that are not differentially methylated to investigate whether they are enriched for genes that would be considered dosage-sensitive.

Table 7.8: Genes associated with most significant DMRs between low and high susceptibility trees. DMRs not associated with genes are not shown here.

DMR	Gene ID	Gene location	Gene Function	Mean methylation level		p-value
				High Sus.	Low Sus.	
CG						
Contig2609: 64538-64958	137200	Contig2609: 64196-72467	mechanosensitive ion channel domain-containing protein	0.65	0.002	1.5e-29
Contig3629: 24849-25184	197890	Contig3629: 21463-28194	None	2.2e-08	0.746	8.7e-22
Contig324: 289772-289959	176630	Contig324: 289086-296300	vacuolar sorting protein 4b	0.419	0.0026	2.3e-21
Contig2911: 138275-138754	157480	Contig2911: 138036-139254	Senescence regulator S40	0.468	0.0031	2.8e-17
Contig1417: 73501-73781	043910	Contig1417: 70658-74130	None	0.441	0.0059	2.4e-16
Contig4146: 18321-18487	224530	Contig4146: 9934-20250	None	0.529	0.898	1.9e-15
Contig1807: 87341-87737	075340	Contig1807: 79037-92335	None	0.97144	0.60362	2.7e-15
Contig6324: 7968-8375	310520	Contig6324: 7530-8441	mads-box protein	0.402	0.096	4.4e-15
Contig185: 21549-21735	078870	Contig185: 10869-21735	None	0.74566	0.96215	7.3e-15
Contig66908: 5291-5613	320530	Contig66908: 3616-6025	probable protein phosphatase 2c 8-like	0.459	0.0058	4.8e-14
CHG						
Contig1991: 59224-59868	090210	Contig1991: 56127-85591	Serine hydroxymethyltransferase	0.216	0.824	1.9e-25
Contig1417: 73461-73861	043910	Contig1417: 70658-74130	None	0.421	0.0011	2.4e-25
Contig710: 99778-100556	332070	Contig710: 99419-100605	germin-like protein 2-1-like	0.857	0.381	2.7e-24
Contig2279: 41279-41889	112410	Contig2279: 41066-42067	blue copper protein	0.564	0.017	5.6e-19
Contig2609: 64573-64977	137200	Contig2609: 64196-72467	mechanosensitive ion channel domain-containing protein	0.649	0.009	3.1e-16
Contig3629: 24832-25193	197890	Contig3629: 21463-28194	None	5.1e-08	0.558	4.9e-15
Contig71553: 14558-15560	333290	Contig71553: 11322-20261	0-methyltransferase	1.07e-07	0.278	8.6e-15
Contig523: 53646-54030	268290	Contig523: 51706-57839	None	0.807	0.158	9.1e-13
Contig2118: 69725-70643	100650	Contig2118: 62961-72522	la protein 1	0.779	0.248	3.7e-12
Contig32236: 422-788	175600	Contig32236: 1-1169	None	0.392	0.040	4.9e-12
CHH						
Contig1183: 57475-57539	021870	Contig1183: 57196-64032	None	0.912	0.657	2.9e-26
Contig20: 381128-381219	091230	Contig20: 379336-383904	actin associated protein	0.893	0.745	3.3e-21
Contig85: 162243-162327	367270	Contig85: 158536-169445	lipase, class 3	0.953	0.766	3.5e-20
Contig1829: 17191-17265	077110	Contig1829: 6282-34261	cytochrome p450	0.952	0.735	1.1e-19
Contig2210: 13175-13268	107820	Contig2210: 11718-16741	None	0.550	0.331	3.2e-18
Contig958: 48695-48770	392290	Contig958: 38756-53536	heat shock protein 70	0.963	0.803	1.2e-17
Contig641: 10536-10609	313160	Contig641: 4064-20283	protein transport protein	0.896	0.689	2.8e-17
Contig1409: 33675-33753	043220	Contig1409: 29218-37473	proline-rich family protein	0.934	0.781	4.3e-17
Contig8217: 49802-49942	360250	Contig8217: 49285-54385	None	0.739	0.54059	5.4e-17
Contig3661: 49737-49812	199700	Contig3661: 48853-65255	calmodulin-like protein	0.581	0.134	1.1e-16

Chapter 8

Conclusions and further research

8.1 Summary of results

This thesis has covered many aspects of the ash tree genome and epigenome, which would not have been possible without first sequencing and assembling a reference genome sequence. Very little genetic data were available for the ash tree prior to this project; mainly ESTs and some microsatellite markers used in population structure studies. I have assembled the nuclear, mitochondrial, and plastid genomes to a sufficient level to allow further meaningful analyses to be carried out. Though the genomes are all still in a draft, fragmented format, gene annotations have been performed and used in other parts of this thesis, and indeed by other researchers outside of the project. For example, the cDNA assembly was used in a study on associative transcriptomics of ADB susceptibility, as a reference for mapping cDNA reads [Harper et al. 2016], and by a group in the US to aid an RNA-seq study into expression changes in response to environmental stresses in *F. pennsylvanica*[T. Lane, pers. comm.]. It is also currently aiding the assembly of genomes from various *Fraxinus* species in a genus-wide study of ADB and EAB susceptibility [L. Kelly, pers.comm.]

I used the genome sequence as a reference to map reads from 38 European ash trees originating throughout the species' range. Using a set of nearly 400,000 polymorphic positions, I find little evidence of population structure in the group using three different methods, in agreement with previous research using nuclear genome markers. In contrast, research using chloroplast markers finds highly structured populations that follow patterns depicting the colonisation route of ash from glacial refugia, northwards into central Europe. This difference is likely due to the maternal inheritance of plastid DNA, which only allows it to be passed through seeds and therefore is not dispersed as far as pollen and its nuclear DNA. Using the linkage information between the polymorphic loci, I find that the effective population size of European ash has been decreasing for the past 4000 generations (approximately 60,000-80,000 years). Using a TMRCA model, I estimate that the effective population size peaked approximately 20 million years ago during the warm Miocene and subsequently fell during the cooling of the Miocene and Pliocene. Glaciation events during the Quaternary period caused many temperate species to shift their ranges into refugia until climate warming allowed re-colonisation into central Europe.

Using an analysis of Ks values (synonymous substitutions per synonymous site), I provide

evidence for two recent WGD events in the ash lineage. Comparing the Ks distribution of ash with five other species (olive, monkey flower, bladderwort, coffee, tomato and grape), I show that the most recent WGD event is likely shared with olive, and could be common to all in the Oleaceae family. The older of the two WGD events could be shared with monkey flower and bladderwort, for which a shared WGD event has already been documented, and could be common to all in the Lamiales order. Further species in the respective groups could be tested to see if these hypotheses hold true, though this would require whole transcriptome sequencing. I also find that many of the homeologs in ash are differentially methylated, meaning that pairs of homeologs retained from the WGD events could be unequally silenced. Gene silencing is commonly employed as a dosage compensation mechanism post-polyploidisation, and often one paralog is preferentially expressed over another.

In addition to the unequal silencing of homeologs, I find methylation levels in ash comparable to that of other plant species using whole genome bisulphite sequencing; especially those of tree species poplar and birch, as well as fellow Asterid tomato. In keeping with previous research, I find that transposable elements are heavily methylated along their whole length in the ash genome compared to non-mobile genes. Housekeeping genes, particularly those with functions related to photosynthesis and respiration, have a high density of unmethylated sites. I find evidence of species- and genotype-specific methylation patterns, but still significant variability in the ash epigenome between genetically identical individuals. The epigenome is much more dynamic than the genome, as the epimutation rate is several orders of magnitude higher than the sequence mutation rate [Becker et al. 2011]. This epigenetic variability could help to generate phenotypic plasticity in the population in order to be able to adapt quickly to changing environments.

The genome sequence of the ash tree has been imperative to all further genetic research carried out so far on European ash, and will continue to be in the future. The usefulness of the gene annotation and genome sequence for researchers outside of the immediate project group, demonstrates the benefit of the sequence and also of the open access policy to the research community and ultimately to the conservation of the ash tree.

8.2 Future research using the ash genome

Many of the results and data from this project can be used by other researchers to improve understanding about aspects of the ash genome or population as a whole. The genome sequence, as mentioned previously, is a fundamental foundation to any future genetic research on the ash tree. It allows the development of primers and markers to genotype specific regions of the genome. This could be used to gain more data on the phylogeographic structure of the ash population or develop markers for association studies with certain traits of interest. Indeed, a set of polymorphic microsatellite markers were developed from the BATG-0.5 reference genome and diversity panel of 38 European trees (work performed by Gemma Worswick and published in Sollars et al. 2017). As a reference sequence, DNA reads from any *F. excelsior* tree can be mapped and compared, as well as other closely related species in the *Fraxinus* genus also. This opens the door for comparative genomics and phylogenomics studies which will benefit from the functional gene annotations available. RNA-seq data

can also be mapped to the reference, and therefore research into gene expression changes over different conditions can be carried out easily, again benefiting from the functional gene annotation.

The genome sequence is still in a draft, fragmented form, however. Though a genome may never be considered truly 'complete', the ash assembly can definitely be improved upon. A largely contiguous assembly with scaffolds anchored onto chromosomes would first require a linkage map, for which a mapping population of trees has not been generated. The greatest improvements possible within a short time frame may come from long-read NGS technologies such as PacBio or Oxford Nanopore. Long reads of several thousand base pairs (often tens of thousands) would allow scaffolds to be linked together, over-assembled repeat regions to be separated, and gaps to be filled. In addition, optical mapping technology such as BioNano Irys could also improve contiguity by helping to order and join scaffolds. A more contiguous assembly improves many further analyses by containing additional sequence (e.g., filling in gaps) and by reducing erroneous duplications in the assembly. DNA reads are therefore more likely to map with less fragmented scaffolds, especially those using long range mate pair libraries that span large distances. Haplotype construction becomes possible with a contiguous assembly also, accompanied by uniform read coverage and/or long reads, as the linkage group information over thousands to millions of base pairs can be discerned.

In this thesis, I examined the population structure in European ash using 38 trees originating from across the native range and nearly 400,000 genomic markers. I found little evidence of population structure or geographical patterns. However, I believe this study could be improved upon by sampling more trees from across the range. Firstly, multiple individuals from each location / population should be sampled so as to ensure that the genotypes sampled are representative of the location. Using only one individual from each location, as in this study, increases the chance of stochastic variation influencing the results, i.e. we may have sampled rare genotypes by chance in some locations which would not be representative of the gene pool. Secondly, the trees used this study had all been selected for a trial based on silvicultural value. Silvicultural value is a very broad trait that could include many phenotypes such as wood quality, straight trunks or fast growth that would all likely be controlled by very many loci. It is therefore very unlikely that this selection of 38 trees would have certain alleles enriched at specific loci. However, there still remains an element of non-randomness about the sample selection that could be overcome if trees were selected at random from each location. These would then encompass more of the genetic variation present in natural populations. Lastly, there were some areas of the *F. excelsior* natural range that were missed in this study, particularly those identified as probable refugia during the Last Glacial Maximum (LGM) such as the Iberian, Italian and Balkan peninsulas, and the Carpathian and Apennine mountains. Trees in these areas could possess additional genetic variation to the 38 trees sampled and could therefore change the conclusions about population structure.

One of the main aims of the British Ash Tree Genome Project is to provide the reference sequence and additional data to researchers studying susceptibility to ash dieback disease. Some expression markers that appear to predict susceptibility have already been found by

collaborators [Harper et al. 2016; Sollars et al. 2017]. However, the heritability of these markers is yet to be determined; if the heritability is low, selection for these expression markers will take many generations to improve the susceptibility phenotype. Genomic sequence markers greatly speed up the selection process, especially so in an organism with such a long time to reproduction. Other researchers aim to identify resistant trees among natural populations. Both of these resources could prove valuable if breeding trials for low susceptibility are initiated.

The candidate genes for ADB susceptibility identified in this thesis using bisulphite sequencing are good starting points as epigenetic markers, but they were found using such a small sample size that the probability of methylation differences between the groups due to chance is very high leading to false positive results. On the other hand, the statistical tests used may not have had sufficient power to be able to reject the null hypothesis such a small sample size, leading to false negative results. Further studies could repeat the methylation experiment on a larger group of samples to minimise chance effects and also incorporate a diverse genetic background. The panels of trees used by our collaborators in Harper et al. (2016) and Sollars et al. (2017) could be of use here as susceptibility phenotype data have already been recorded, some over several years, saving much work in measuring disease damage in a new set of trees. It would also be useful to obtain RNA-seq data for the trees studied in order to tie in any methylation patterns identified with changes in gene expression. If methylation or expression changes are found and verified in a testing panel of samples, these could be developed into markers that could be used to predict susceptibility of seedlings in breeding trials. The prediction of phenotypic traits speeds up the selection process as researchers can perform controlled crosses of parent trees, that are predicted to have the greatest improvement in the next generation, thereby saving time, resources and money.

Other genes that could be involved in low ADB susceptibility could be the Resistance (R) genes. R genes are a group of genes encoding NBS-LRR (Nucleotide-Binding Site Leucine Rich Repeat) domains that evolve quickly using recombination in order to diversify. The full complement of R gene sequences in a population possess large amounts of variation to adapt quickly to pest or pathogen infection and therefore, could be involved in the ADB susceptibility genotype. R genes have not been studied specifically in the ash tree genome yet, though there are 45 transcripts already annotated with an 'NBS-LRR' domain. As it is likely that there are more R genes in the ash genome, these might require a more manual identification and annotation; the Plant Resistance Genome Database (http://prgdb.crg.eu) could be useful for this. For example, there are 259 genes in the ash genome that have a BLAST hit to a gene from the R gene database with over 50% identity. Furthermore, enrichment methods for sequencing R genes have been developed. RenSeq (R gene enrichment and sequencing) [Jupe et al. 2013] targets NBS-LRR sequences and allows firstly, the annotation of R genes without relying on a reference sequence and secondly, the identification of markers in R genes that segregate with phenotypes such as resistance.

Finally, an experiment in breeding low ADB susceptibility could investigate the hybridisation of *F. excelsior* with other *Fraxinus* species, particularly Asian species such as *F.*

mandshurica or *F. chinensis*. These species are naturally resistant to the fungus due to co-evolution. Hybridisation as a method for introducing resistant alleles into natural populations has been studied widely in the American chestnut. The tree has been hybridised with Chinese chestnut to introduce resistance to chestnut blight, markers associated with resistance have been developed, and the potential of genetic modification has also been investigated [Thompson 2012; Hebard et al. 2014]. A hybridisation study on ash would raise interesting research questions. Firstly, whether the progeny do indeed display signs of low susceptibility; an intermediate phenotype would be expected in the F1 generation, if the susceptibility trait is determined by many loci. However, the phenotype may show more segregation in the F2 and following generations. Secondly, whether one species' alleles are expressed preferentially over the other; this could be investigated using expression data such as RNA-seq, as has been performed for homoeologous loci in hexaploid bread wheat [Leach et al. 2014], or using Bisulphite-seq to study Allele Specific Methylation (ASM). Indeed, studies of ASM have been carried out in *Arabidopsis* [Peng & Ecker 2012; Chen et al. 2014]. Detection of ASM requires SNPs to be present within a read length of a methylated cytosine so that the methylation level for each allele of the SNP can be calculated. As sequence mutations are less frequent than epimutations, there may be regions of the epigenome that lack SNPs, however. Polymorphism data from the European ash population study (described in Chapter 6) as well as data from the genus-wide study that includes *F. mandshurica* could be used to identify regions with high genetic diversity. These regions would be excellent candidates for studying ASM in hybrid trees.

There is now much scope for future genetic research on *Fraxinus excelsior*, to which the reference genome sequence and annotation will provide the foundations. A valuable and urgent use of the sequence will be in research aiming to breed a population of trees with low susceptibility to ADB. The achievement of this goal will be accelerated by the use of genomic markers in selecting low susceptibility progeny from breeding trials. The techniques used for many years to improve traits in other plants can also be successful in this tree species. Markers can also be useful in ensuring genetic diversity is maintained at common polymorphic loci. Genetic diversity is imperative to maintaining sufficient variation, in order to adapt to future environmental and biological stresses, especially in an already vulnerable population. Coupled with the low susceptibility found in natural ash stands, there is certainly reason to be hopeful that populations of *Fraxinus excelsior* will not disappear completely from Europe.

Bibliogaphy

Akimoto, K., Katakami, H., Kim, H. J., Ogawa, E., Sano, C. M., Wada, Y. & Sano, H. (2007) Epigenetic inheritance in rice plants. Annals of Botany, 100(2), 205-217.

Al-Dous, E. K., George, B., Al-Mahmoud, M. E., et al. (2011) De novo genome sequencing and comparative genomics of date palm (*Phoenix dactylifera*). Nature Biotechnology, 29(6), 521-U84.

Al-Mssallem, I. S., Hu, S. N., Zhang, X. W., et al. (2013) Genome sequence of the date palm *Phoenix dactylifera* L. Nature Communications, 4.

Anagnostakis, S. L. (1987) CHESTNUT BLIGHT - THE CLASSICAL PROBLEM OF AN INTRODUCED PATHOGEN. Mycologia, 79(1), 23-37.

Argout, X., Salse, J., Aury, J. M., et al. (2011) The genome of *Theobroma cacao*. Nature Genetics, 43(2), 101-108.

Ausin, I. Feng, S., Yu, C. et al. (2016) DNA methylome of the 20-gigabase Norway spruce genome. PNAS, 113(50), E8106-E8113.

Bacles, C. F. E., Burczyk, J., Lowe, A. J. & Ennos, R. A. (2005) Historical and contemporary mating patterns in remnant populations of the forest tree *Fraxinus excelsior* L. Evolution, 59(5), 979-990.

Bacles, C. F. E. & Ennos, R. A. (2008) Paternity analysis of pollen-mediated gene flow for *Fraxinus excelsior* L. in a chronically fragmented landscape. Heredity, 101(4), 368-380.

Ballester, A. R., Lafuente, M. T., de Vos, R. C. H., Bovy, A. G. & Gonzalez-Candelas, L. (2013) Citrus phenylpropanoids and defence against pathogens. Part I: Metabolic profiling in elicited fruits. Food Chemistry, 136(1), 178-185.

Bao, Y. & Xu, Q. (2015) Extensive reprogramming of cytosine methylation in *Oryza* allotetraploids. Genes & Genomics, 37(6), 517-524.

Baral, H., Queloz, V., and Hosoya, T. (2014). *Hymenoscyphus Fraxineus*, the Correct Scientific Name for the Fungus Causing Ash Dieback in Europe. IMA Fungus 5: 7980.

Barbato, M., Orozco-terWengel, P., Tapio, M., Bruford, M.W. (2015). SNeP: a tool to estimate trends in recent effective population size trajectories using genome-wide SNP data. Frontiers in Genetics, 6: 109.

Barr, C. M., Neiman, M. & Taylor, D. R. (2005) Inheritance and recombination of mitochondrial genomes in plants, fungi and animals. New Phytologist, 168(1), 39-50.

Beatty, G. E., Brown, J. A., Cassidy, E. M., Finlay, C. M. V., McKendrick, L., Montgomery, W. I., Reid, N., Tosh, D. G. & Provan, J. (2015) Lack of genetic structure and evidence for long-distance dispersal in ash (*Fraxinus excelsior*) populations under threat from an emergent fungal pathogen: implications for restorative planting. Tree Genetics & Genomes, 11(3).

Becker, C., Hagmann, J., Muller, J., Koenig, D., Stegle, O., Borgwardt, K. & Weigel, D. (2011) Spontaneous epigenetic variation in the *Arabidopsis thaliana* methylome. Nature, 480(7376), 245-U127.

Beilstein M.A., Nagalingum N.S., Clements M.D., Manchester S.R., Mathews S. (2010) Dated molecular phylogenies indicate a Miocene origin for *Arabidopsis thaliana*. Proc Natl Acad Sci USA, 107:18724-18728.

Bilichak, A. & Kovalchuk, I. (2016) Transgenerational response to stress in plants and its application for breeding. Journal of Experimental Botany, 67(7), 2081-2092.

Binggeli, P, & Power, J. (1991). Gender variation in ash (*Fraxinus excelsior* L.). Miscellaneous Notes & Reports in Natural History, Ecology, Conservation and Resources Management. Accessed from http://www.mikepalmer.co.uk/woodyplantecology/docs/MNR-ashgender.pdf on 20 November 2016.

Birchler, J. A. & Veitia, R. A. (2007) The gene balance hypothesis: From classical genetics to modern genomics. Plant Cell, 19(2), 395-402.

Birol, I., Raymond, A., Jackman, S. D., et al. (2013) Assembling the 20 Gb white spruce (*Picea glauca*) genome from whole-genome shotgun sequencing data. Bioinformatics, 29(12), 1492-1497.

Blanc, G. & Wolfe, K. H. (2004) Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. Plant Cell, 16(7), 1667-1678.

Bock, C. (2012) Analysing and interpreting DNA methylation data. Nature Reviews Genetics, 13(10), 705-719.

Bock, R. & Timmis, J.N. (2008). Reconstructing evolution: gene transfer from plastids to the nucleus. BioEssays, 30(6), 556-566.

Boetzer, M., Henkel, C. V., Jansen, H. J., Butler, D. & Pirovano, W. (2011) Scaffolding pre-assembled contigs using SSPACE. Bioinformatics, 27(4), 578-579.

Boetzer, M. & Pirovano, W. (2012) Toward almost closed genomes with GapFiller. Genome Biology, 13(6).

Bomblies, K. & Madlung, A. (2014) Polyploidy in the *Arabidopsis* genus. Chromosome Research, 22(2), 117-134.

Bradshaw, H. D., Ceulemans, R., Davis, J. & Stettler, R. (2000) Emerging model systems in plant biology: Poplar (*Populus*) as a model forest tree. Journal of Plant Growth Regulation, 19(3), 306-313.

Bradshaw, H. D. & Stettler, R. F. (1995) MOLECULAR-GENETICS OF GROWTH AND DEVELOPMENT IN POPULUS .4. MAPPING QTLS WITH LARGE EFFECTS ON GROWTH, FORM, AND PHENOLOGY TRAITS IN A FOREST TREE. Genetics, 139(2), 963-973.

Brenchley, R., Spannagl, M., Pfeifer, M., et al. (2012) Analysis of the breadwheat genome using whole-genome shotgun sequencing. Nature, 491(7426), 705-710.

Brewer, S., Cheddadi, R., de Beaulieu, J. L., & Reille, M.(2002) The spread of deciduous *Quercus* throughout Europe since the last glacial period. Forest Ecology and Management, 156(1-3), 27-48.

Brown, G. R., Gill, G. P., Kuntz, R. J., Langley, C. H. & Neale, D. B. (2004) Nucleotide diversity and linkage disequilibrium in loblolly pine. Proceedings of the National Academy of Sciences of the United States of America, 101(42), 15255-15260.

Brownfield, L. & Kohler, C (2011). Unreduced gamete formation in plants: mechanisms and prospects. Journal of Experimental Botany, 62(5), 1659-1668.

Buggs, R. J. A., Elliott, N. M., Zhang, L. J., Koh, J., Viccini, L. F., Soltis, D. E. & Soltis, P. S. (2010) Tissue-specific silencing of homoeologs in natural populations of the recent allopolyploid *Tragopogon mirus*. New Phytologist, 186(1), 175-183.

Buschiazzo, E., Ritland, C., Bohlmann, J. z & Ritland, K. (2012) Slow but not low: genomic comparisons reveal slower evolutionary rate and higher dN/dS in conifers compared to angiosperms. BMC Evolutionary Biology, 12.

Campoy, J. A., Lerigoleur-Balsemin, E., Christmann, H., Beauvieux, R., Girollet, N., Quero-Garcia, J., Dirlewanger, E. & Barreneche, T. (2016) Genetic diversity, linkage disequilibrium, population structure and construction of a core collection of *Prunus avium* L. landraces and bred cultivars. BMC Plant Biology, 16.

Carlson, C.H., Gouker, F.E., DiFazio, S., Zhou, R., and Smart, L. 2016. High-resolution mapping of biomass-related traits in shrub willow (*Salix purpurea* L.). In Plant and Animal Genome XXIV Conference. Plant and Animal Genome. https://pag.confex.com/pag/xxiv/webprogram/Paper21612.html.

Carlson, J.E. 2014. The chestnut genome project. In Plant and Animal Genome XXII Conference. Plant and Animal Genome. https://pag.confex.com/pag/xxii/webprogram/Paper9777.html.

Chen, S. X., He, H. & Deng, X. W. (2014) Allele-specific DNA methylation analyses associated with siRNAs in Arabidopsis hybrids. Science China-Life Sciences, 57(5), 519-525.

Chen, Z. J. (2007) Genetic and epigenetic mechanisms for gene expression and phenotypic variation in plant polyploids. Annual Review of Plant Biology, 58, 377-406.

Cheng, F., Wu, J., Fang, L., Sun, S. L., Liu, B., Lin, K., Bonnema, G. & Wang, X. W. (2012) Biased Gene Fractionation and Dominant Gene Expression among the Subgenomes of *Brassica rapa*. Plos One, 7(5).

Cormier, Z. (2012). UK unveils plan to fight deadly ash disease. Nature News doi:10.1038/nature.2012.11790.

Cruz, F., Julca, I., Gomez-Garrido, J., et al. (2016) Genome sequence of the olive tree, *Olea europaea*. Gigascience, 5.

Cuenca, A., Petersen, G. & Seberg, O. (2013) The Complete Sequence of the Mitochondrial Genome of *Butomus umbellatus* - A Member of an Early Branching Lineage of Monocotyledons. Plos One, 8(4).

Dai, X. G., Hu, Q. J., Cai, Q. L., et al. (2014) The willow genome and divergent evolution from poplar after the common genome duplication. Cell Research, 24(10), 1274-1277.

Davila, J. I., Arrieta-Montiel, M. P., Wamboldt, Y., Cao, J., Hagmann, J., Shedge, V., Xu, Y. Z., Weigel, D. & Mackenzie, S. A. (2011) Double-strand break repair processes drive evolution of the mitochondrial genome in Arabidopsis. BMC Biology, 9.

DEFRA (2013). Chalara in Ash Trees: A framework for assessing ecosystem impacts and appraising options. Downloaded from www.gov.uk.defra in January 2014.

Denman, S. and Webber, J. (2013). Chalara Dieback of Ash. Last accessed from www.forestry.gov.uk/forestresearch in November 2016.

Denoeud, F., Carretero-Paulet, L., Dereeper, A., et al. (2014) The coffee genome provides insight into the convergent evolution of caffeine biosynthesis. Science, 345(6201), 1181-1184.

Dobrowolska, D., Hein, S., Oosterbaan, A., Wagner, S., Clark, J. & Skovsgaard, J. P. (2011) A review of European ash (*Fraxinus excelsior* L.): implications for silviculture. Forestry, 84(2), 133-148.

Doyle, J. J. and Doyle, J. L. (1987). A rapid DNA isolation procedure for small quantities of fresh leaf tissue. Phytochemical Bulletin, 19: 1115.

Earl, D. A. & Vonholdt, B. M. (2012) STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. Conservation Genetics Resources, 4(2), 359-361.

Elo, A., Lyznik, A., Gonzalez, D.O., Kachman, S.D., Mackenzie, S.A. (2003) Nuclear Genes That Encode Mitochondrial Proteins for DNA and RNA Metabolism Are Clustered in the Arabidopsis Genome. Plant Cell, 15(7), 1619-1631.

English, A. C., Richards, S., Han, Y., et al. (2012) Mind the Gap: Upgrading Genomes with Pacific Biosciences RS Long-Read Sequencing Technology. Plos One, 7(11).

EUFORGEN (2009). Distribution map of Common ash (*Fraxinus excelsior*) EUFORGEN 2009, www.euforgen.org.

Evanno, G., Regnaut, S. & Goudet, J. (2005) Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. Molecular Ecology, 14(8), 2611-2620.

Excoffier, L., Dupanloup, I., Huerta-Sanchez, E., Sousa, V. C. & Foll, M. (2013) Robust Demographic Inference from Genomic and SNP Data. Plos Genetics, 9(10).

Fabio, E. S., Volk, T. A., Miller, R. O., et al. (2016), Genotype environment interaction analysis of North American shrub willow yield trials confirms superior performance of triploid hybrids. GCB Bioenergy. doi:10.1111/gcbb.12344.

Fan, D., Liu, T. T., Li, C. F., Jiao, B., Li, S., Hou, Y. S. & Luo, K. M. (2015) Efficient CRISPR/Cas9-mediated Targeted Mutagenesis in Populus in the First Generation. Scientific Reports, 5.

Fang, G. C., Blackmon, B. P., Staton, M. E., et al. (2013) A physical map of the Chinese chestnut (*Castanea mollissima*) genome and its integration with the genetic map. Tree Genetics & Genomes, 9(2), 525-537.

Fedoroff, N. V. (2012) PRESIDENTIAL ADDRESS Transposable Elements, Epigenetics, and Genome Evolution. Science, 338(6108), 758-767.

Fones, H.N., Mardon, C., & Gurr, S.J. (2016) A role for the asexual spores in infection of *Fraxinus excelsior* by the ash-dieback fungus *Hymenoscyphus fraxineus*. Scientific Reports, 6, Article number: 34638.

Freeling, M. (2009) Bias in Plant Gene Content Following Different Sorts of Duplication: Tandem, Whole-Genome, Segmental, or by Transposition. Annual Review of Plant Biology, 60, 433-453.

Freeling, M., Woodhouse, M. R., Subramaniam, S., Turco, G., Lisch, D. & Schnable, J. C. (2012) Fractionation mutagenesis and similar consequences of mechanisms removing dispensable or less-expressed DNA in plants. Current Opinion in Plant Biology, 15(2), 131-139.

Fussi, B., Lexer, C. & Heinze, B. (2010) Phylogeography of *Populus alba* (L.) and *Populus tremula* (L.) in Central Europe: secondary contact and hybridisation during recolonisation from disconnected refugia. Tree Genetics & Genomes, 6(3), 439-450.

Gao, M., Huang, Q., Chu, Y., Ding, C., Zhang B. & Su, X. (2014) Analysis of the leaf methylomes of parents and their hybrids provides new insight into hybrid vigor in *Populus deltoides*. BMC Genomics, 15(Suppl1): S8.

Gerard, P. R., Temunovic, M., Sannier, J., Bertolino, P., Dufour, J., Frascaria-Lacoste, N. & Fernandez-Manjarres, J. F. (2013) Chilled but not frosty: understanding the role of climate in the hybridization between the Mediterranean *Fraxinus angustifolia* Vahl and the temperate *Fraxinus excelsior* L. (Oleaceae) ash trees. Journal of Biogeography, 40(5), 835-846.

Glenz, C., Schlaepfer, R., Iorgulescu, I. & Kienast, F. (2006) Flooding tolerance of Central European tree and shrub species. Forest Ecology and Management, 235(1-3), 1-13.

Gouker, F.E., Zhou, R., Evans, L., DiFazio, S., Bubner, B., Zander, M., & Smart, L. (2016). Genotypic-phenotypic variation and marker-based heritability estimates of a shrub willow (*Salix purpurea*) association population. In Plant and Animal Genome XXIV Conference. Plant and Animal Genome. https://pag.confex.com/pag/xxiv/webprogram/Paper19730.html.

Greiner, S. & Bock, R. (2013). Tuning a mnage trois: Co-evolution and co-adaptation of nuclear and organellar genomes in plants. BioEssays, 35(4), 354-365.

Gross, A., Holdenrieder, O., Pautasso, M., Queloz, V. & Sieber, T. N. (2014) *Hymenoscyphus pseudoalbidus*, the causal agent of European ash dieback. Molecular Plant Pathology, 15(1), 5-21.

Gross, A. & Sieber, T. N. (2016) Virulence of *Hymenoscyphus albidus* and native and introduced *Hymenoscyphus fraxineus* on *Fraxinus excelsior* and *Fraxinus pennsylvanica*. Plant Pathology, 65(4), 655-663.

Gugger, P. F., Fitz-Gibbon, S., Pellegrini, M. & Sork, V. L. (2016) Species-wide patterns of DNA methylation variation in *Quercus lobata* and their association with climate gradients. Molecular Ecology, 25(8), 1665-1680.

Gutenkunst, R. N., Hernandez, R. D., Williamson, S. H. & Bustamante, C. D. (2009) Inferring the Joint Demographic History of Multiple Populations from Multidimensional SNP Frequency Data. Plos Genetics, 5(10).

Hackl, T., Hedrich, R., Schultz, J. & Forster, F. (2014) proovread: large-scale high-accuracy PacBio correction through iterative short read consensus. Bioinformatics, 30(21), 3004-3011.

Harper, A. L., McKinney, L. V., Nielsen, L. R., et al. (2016) Molecular markers for tolerance of European ash (*Fraxinus excelsior*) to dieback disease identified using Associative Transcriptomics. Scientific Reports, 6.

Hauben, M., Haesendonckx, B., Standaert, E., et al. (2009) Energy use efficiency is characterized by an epigenetic component that can be directed through artificial selection to increase yield. Proceedings of the National Academy of Sciences of the United States of America, 106(47), 20109-20114.

Hebard, F. V., Islam-Faridi, N., Staton, M.E., and Georgi, L. (2014). Biotechnology of trees: chestnut. Tree Biotechnology. CRC Press, 1.

Hegarty, M. J. (2012) Invasion of the hybrids. Molecular Ecology, 21(19), 4669-4671.

Helm, D. (2015). Natural Capital: Valuing the Planet. Yale University Press.

Henn, B. M., Cavalli-Sforza, L. L. & Feldman, M. W. (2012) The great human expansion. Proceedings of the National Academy of Sciences of the United States of America, 109(44), 17758-17764.

Herms, D. A. & McCullough, D. G. (2014) Emerald Ash Borer Invasion of North America: History, Biology, Ecology, Impacts, and Management. Annual Review of Entomology, Vol 59, 2014, 59, 13-30.

Heuertz, M., Carnevale, S., Fineschi, S., Sebastiani, F., Hausman, J. F., Paule, L. & Vendramin, G. G. (2006) Chloroplast DNA phylogeography of European ashes, *Fraxinus* sp (Oleaceae): roles of hybridization and life history traits. Molecular Ecology, 15(8), 2131-2140.

Heuertz, M., Hausman, J. F., Hardy, O. J., Vendramin, G. G., Frascaria-Lacoste, N. & Vekemans, X. (2004a) Nuclear microsatellites reveal contrasting patterns of genetic structure between western and southeastern European populations of the common ash (*Fraxinus excelsior* L.). Evolution, 58(5), 976-988.

Heuertz, M., Fineschi, S., Anzidei, M., Pastorelli, R., Salvini, D., Paule, L., Frascaria-Lacoste, N., Hardy, O. J., Vekemans, X. & Vendramin, G. G. (2004b) Chloroplast DNA variation and postglacial recolonization of common ash (*Fraxinus excelsior* L.) in Europe. Molecular Ecology, 13(11), 3437-3452.

Hollister, J. D. (2015) Polyploidy: adaptation to the genomic environment. New Phytologist, 205(3), 1034-1039.

Howe, K. & Wood, J. M. D. (2015) Using optical mapping data for the improvement of vertebrate genome assemblies. Gigascience, 4.

Ibarra-Laclette, E., Lyons, E., Hernandez-Guzman, G., et al. (2013) Architecture and evolution of a minute plant genome. Nature, 498(7452), 94.

Ingvarsson, P.K., Hvidsten, T.R., & Street, N.R. (2016). Towards integration of population and comparative genomics in forest trees. New Phytologist, 212(2), 338-344.

Ioos, R., Kowalski, T., Husson, C. & Holdenrieder, O. (2009) Rapid in planta detection of *Chalara fraxinea* by a real-time PCR assay using a dual-labelled probe. European Journal of Plant Pathology, 125(2), 329-335.

Jaillon, O., Aury, J. M., Noel, B., et al. (2007) The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature, 449(7161), 463.

Jiao, Y. N., Wickett, N. J., Ayyampalayam, S., et al. (2011) Ancestral polyploidy in seed plants and angiosperms. Nature, 473(7345), 97-U113.

Juhling, F., Kretzmer, H., Bernhart, S. H., Otto, C., Stadler, P. F. & Hoffmann, S. (2016) metilene: fast and sensitive calling of differentially methylated regions from bisulfite sequencing data. Genome Research, 26(2), 256-262.

Jupe, F., Witek, K., Verweij, W., et al. (2013) Resistance gene enrichment sequencing (RenSeq) enables reannotation of the NB-LRR gene family from sequenced plant genomes and rapid mapping of resistance loci in segregating populations. Plant Journal, 76(3), 530-544.

Kaul, S., Koo, H. L., Jenkins, J., et al. (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature, 408(6814), 796-815.

Kerr, G. & Cahalan, C. (2004) A review of site factors affecting the early growth of ash (*Fraxinus excelsior* L.). Forest Ecology and Management, 188(1-3), 225-234.

Kim, K. J. & Jansen, R. K. (1995) NDHF SEQUENCE EVOLUTION AND THE MAJOR CLADES IN THE SUNFLOWER FAMILY. Proceedings of the National Academy of Sciences of the United States of America, 92(22), 10379-10383.

Kirisits, T., Matlakova, M., Mottinger-Kroupa, S., Halmschlager, E. & Lakatos, F. (2010) *Chalara fraxinea* associated with dieback of narrow-leafed ash (*Fraxinus angustifolia*). Plant Pathology, 59(2), 411-411.

Kirisits, T. & Schwanda, K. (2015) First definite report of natural infection of *Fraxinus ornus* by *Hymenoscyphus fraxineus*. Forest Pathology, 45(5), 430-432.

Kjaer, E. D., McKinney, L. V., Nielsen, L. R., Hansen, L. N. & Hansen, J. K. (2012) Adaptive potential of ash (*Fraxinus excelsior*) populations against the novel emerging pathogen *Hymenoscyphus pseudoalbidus*. Evolutionary Applications, 5(3), 219-228.

Kleine, T., Maier, U.G., Leister, D. (2009). DNA transfer from organelles to the nucleus: the idiosyncratic genetics of endosymbiosis. Annual Review of Plant Biology, 60, 115-138.

Konig, S., Feussner, K., Kaever, A., Landesfeind, M., Thurow, C., Karlovsky, P., Gatz, C., Polle, A. & Feussner, I. (2014) Soluble phenylpropanoids are involved in the defense response of Arabidopsis against *Verticillium longisporum*. New Phytologist, 202(3), 823-837.

Kowalski, T. (2006) *Chalara fraxinea* sp nov associated with dieback of ash (*Fraxinus excelsior*) in Poland. Forest Pathology, 36(4), 264-270.

Kowalski, T. & Holdenrieder, O. (2009) The teleomorph of *Chalara fraxinea*, the causal agent of ash dieback. Forest Pathology, 39(5), 304-308.

Krueger, F. & Andrews, S. R. (2011) Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. Bioinformatics, 27(11), 1571-1572.

Ksiazczyk, T., Kovarik, A., Eber, F., Huteau, V., Khaitova, L., Tesarikova, Z., Coriton, O. & Chevre, A. M. (2011) Immediate unidirectional epigenetic reprogramming of NORs occurs independently of rDNA rearrangements in synthetic and natural forms of a polyploid species *Brassica napus*. Chromosoma, 120(6), 557-571.

Kubisiak, T. L., Nelson, C. D., Staton, M. E., et al. (2013) A transcriptome-based genetic map of Chinese chestnut (*Castanea mollissima*) and identification of regions of segmental homology with peach (*Prunus persica*). Tree Genetics & Genomes, 9(2), 557-571.

LaBonte, N., & Woeste, K.E. 2016. Exploring patterns of sequence variation in regions associated with chestnut blight resistance using whole-genome resequencing of Chinese chestnut (*Castanea mollissima*). In Plant and Animal Genome XXIV Conference. Plant and Animal Genome. https://pag.confex.com/pag/xxiv/webprogram/Paper20702.html.

Landolt, J., Gross, A., Holdenreider, O. & Pautasso, M (2016). Ash dieback due to *Hymenoscyphus fraxineus*: what can be learnt from evolutionary ecology? Plant Pathology, 65(7), 1056-1070.

Lane, T., Best, T., Zembower, N. et al. 2016. The green ash transcriptome and identification of genes responding to abiotic and biotic stresses. BMC Genomics, 17:702.

Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biology, 10(3).

Larkin, M. A., Blackshields, G., Brown, N. P., et al. (2007) Clustal W and clustal X version 2.0. Bioinformatics, 23(21), 2947-2948.

Law, J. A. & Jacobsen, S. E. (2010) Establishing, maintaining and modifying DNA methylation patterns in plants and animals. Nature Reviews Genetics, 11(3), 204-220.

Leach, L. J., Belfield, E. J., Jiang, C. F., Brown, C., Mithani, A. & Harberd, N. P. (2014) Patterns of homoeologous gene expression shown by RNA sequencing in hexaploid bread wheat. BMC Genomics, 15.

Levin, D.A. (2011). The timetable for allopolyploidy in flowering plants. Annals of Botany, 112 (7), 1201-1208.

Levy, A. A. & Feldman, M. (2004) Genetic and epigenetic reprogramming of the wheat genome upon allopolyploidization. Biological Journal of the Linnean Society, 82(4), 607-613.

Li, H. & Durbin, R. (2011) Inference of human population history from individual whole-genome sequences. Nature, 475(7357), 493.

Li, Q., Eichten, S. R., Hermanson, P. J. & Springer, N. M. (2014a) Inheritance Patterns and Stability of DNA Methylation Variation in Maize Near-Isogenic Lines. Genetics, 196(3), 667.

Li, X., Zhu, J. D., Hu, F. Y., et al. (2012) Single-base resolution maps of cultivated and wild rice methylomes and regulatory roles of DNA methylation in plant gene expression. BMC Genomics, 13.

Li, X. M., Gao, W. H., Guo, H. L., Zhang, X. L., Fang, D. D. & Lin, Z. X. (2014b) Development of EST-based SNP and InDel markers and their utilization in tetraploid cotton genetic mapping. BMC Genomics, 15.

Li, Y. X. & Li, W. (2009) BSMAP: whole genome bisulfite sequence MAPping program. BMC Bioinformatics, 10.

Liang, D., Zhang, Z. J., Wu, H. L., et al. (2014) Single-base-resolution methylomes of *Populus trichocarpa* reveal the association between DNA methylation and drought stress. BMC Genetics, 15.

Liang, H. P. & Hilu, K. W. (1996) Application of the matK gene sequences to grass systematics. Canadian Journal of Botany-Revue Canadienne De Botanique, 74(1), 125-134.

Liepelt, S., Cheddadi, R., de Beaulieu, J. L., et al. (2009) Postglacial range expansion and its genetic imprints in *Abies alba* (Mill.) - A synthesis from palaeobotanic and genetic data.

Review of Palaeobotany and Palynology, 153(1-2), 139-149.

Lister, R., O'Malley, R. C., Tonti-Filippini, J., Gregory, B. D., Berry, C. C., Millar, A. H. & Ecker, J. R. (2008) Highly integrated single-base resolution maps of the epigenome in Arabidopsis. Cell, 133(3), 523-536.

Lu, M., Krutovsky, K.V., Nelson, C.D., Koralewski, T.E., Byram, T.D., Loopstra, C.A. (2016) Exome genotyping, linkage disequilibrium and population structure in loblolly pine (*Pinus taeda* L.). BMC Genomics, 17, 730.

Luo, R. B., Liu, B. H., Xie, Y. L., et al. (2012) SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. Gigascience, 1.

Lynch, M. & Conery, J. S. (2000) The evolutionary fate and consequences of duplicate genes. Science, 290(5494), 1151-1155.

Madlung, A. (2013) Polyploidy and its effect on evolutionary success: old questions revisited with new tools. Heredity 110, 99104.

Maere, S., De Bodt, S., Raes, J., Casneuf, T., Van Montagu, M., Kuiper, M. & Van de Peer, Y. (2005) Modeling gene and genome duplications in eukaryotes. Proceedings of the National Academy of Sciences of the United States of America, 102(15), 5454-5459.

Magyari, E. K., Veres, D., Wennrich, V., et al. (2014) Vegetation and environmental responses to climate forcing during the Last Glacial Maximum and deglaciation in the East Carpathians: attenuated response to maximum cooling and increased biomass burning. Quaternary Science Reviews, 106, 278-298.

Manning, K., Tor, M., Poole, M., Hong, Y., Thompson, A. J., King, G. J., Giovannoni, J. J. & Seymour, G. B. (2006) A naturally occurring epigenetic mutation in a gene encoding an SBP-box transcription factor inhibits tomato fruit ripening. Nature Genetics, 38(8), 948-952.

Marcais, G. & Kingsford, C. (2011) A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics, 27(6), 764-770.

Marigo, G., Peltier, J. P., Girel, J. & Pautou, G. (2000) Success in the demographic expansion of *Fraxinus excelsior* L. Trees-Structure and Function, 15(1), 1-13.

Martin, J. A., Witzell, J., Blumenstein, K., Rozpedowska, E., Helander, M., Sieber, T. N. & Gil, L. (2013) Resistance to Dutch Elm Disease Reduces Presence of Xylem Endophytic Fungi in Elms (*Ulmus* spp.). Plos One, 8(2).

Martinez-Garcia, P.J., Crepeau, M.W., Puiu, D. et al. (2016). The walnut (*Juglans regia*) genome sequence reveals diversity in genes coding for the biosynthesis of non-structural

polyphenols. The Plant Journal, 87: 507-532.

Mathew, L. S., Spannagl, M., Al-Malki, A., et al. (2014) A first genetic map of date palm (*Phoenix dactylifera*) reveals long-range genome structure conservation in the palms. BMC Genomics, 15.

Mattioni, C., Martin, M. A., Pollegioni, P., Cherubini, M. & Villani, F. (2013) MICROSATELLITE MARKERS REVEAL A STRONG GEOGRAPHICAL STRUCTURE IN EUROPEAN POPULATIONS OF *CASTANEA SATIVA* (FAGACEAE): EVIDENCE FOR MULTIPLE GLACIAL REFUGIA. American Journal of Botany, 100(5), 951-961.

McClintock, B. (1984) THE SIGNIFICANCE OF RESPONSES OF THE GENOME TO CHALLENGE. Science, 226(4676), 792-801.

McKinney, L. V., Nielsen, L. R., Hansen, J. K. & Kjaer, E. D. (2011) Presence of natural genetic resistance in *Fraxinus excelsior* (Oleraceae) to *Chalara fraxinea* (Ascomycota): an emerging infectious disease. Heredity, 106(5), 788-797.

McKinney, L. V., Thomsen, I. M., Kjaer, E. D. & Nielsen, L. R. (2012) Genetic resistance to *Hymenoscyphus pseudoalbidus* limits fungal growth and symptom occurrence in *Fraxinus excelsior*. Forest Pathology, 42(1), 69-74.

Michael, T. P. (2014) Plant genome size variation: bloating and purging DNA. Briefings in Functional Genomics, 13(4), 308-317.

Min, X. J., Butler, G., Storms, R. & Tsang, A. (2005) OrfPredictor: predicting protein-coding regions in EST-derived sequences. Nucleic Acids Research, 33, W677-W680.

Mitchell, R. J., Beaton, J. K., Bellamy, P. E., et al. (2014) Ash dieback in the UK: A review of the ecological and conservation implications and potential management options. Biological Conservation, 175, 95-109.

Morand-Prieur, M. E., Raquin, C., Shykoff, J. A. & Frascaria-Lacoste, N. (2003) Males outcompete hermaphrodites for seed siring success in controlled crosses in the polygamous *Fraxinus excelsior* (Oleaceae). American Journal of Botany, 90(6), 949-953.

Morse, A. M., Peterson, D. G., Islam-Faridi, M. N., et al. (2009) Evolution of Genome Size and Complexity in *Pinus*. Plos One, 4(2).

Munoz-Merida, A., Gonzalez-Plaza, J. J., Canada, A., et al. (2013) De Novo Assembly and Functional Annotation of the Olive (*Olea europaea*) Transcriptome. DNA Research, 20(1), 93-108.

Myers, S., Fefferman, C. & Patterson, N. (2008) Can one learn history from the allelic spectrum? Theoretical Population Biology, 73(3), 342-348.

Neale, D. B. & Kremer, A. (2011) Forest tree genomics: growing resources and applications. Nature Reviews Genetics, 12(2), 111-122.

Neale, D. B., Wegrzyn, J. L., Stevens, K. A., et al. (2014) Decoding the massive genome of loblolly pine using haploid DNA and novel assembly strategies. Genome Biology, 15(3).

Niederhuth, C. E. & Schmitz, R. J. (2014) Covering Your Bases: Inheritance of DNA Methylation in Plant Genomes. Molecular Plant, 7(3), 472-480.

Nielsen, L.R., McKinney, L.V., Hietala, A.M., Kjaer, E.D. (2016) The susceptibility of Asian, European and North American *Fraxinus* species to the ash dieback pathogen *Hymenoscyphus fraxineus* reflects their phylogenetic history. European Journal of Forest Research, 136, 59-73.

Nystedt, B., Street, N. R., Wetterbom, A., et al. (2013) The Norway spruce genome sequence and conifer genome evolution. Nature, 497(7451), 579-584.

Ong-Abdullah, M., Ordway, J. M., Jiang, N., et al. (2015) Loss of Karma transposon methylation underlies the mantled somaclonal variant of oil palm. Nature, 525(7570), 533.

Ossowski, S., Schneeberger, K., Lucas-Lledo, J. I., Warthmann, N., Clark, R. M., Shaw, R. G., Weigel, D. & Lynch, M. (2010) The Rate and Molecular Spectrum of Spontaneous Mutations in *Arabidopsis thaliana*. Science, 327(5961), 92-94.

Parker, J, Tsagkogeorga, G., Cotton, J.A., Liu, Y., Provero, P., Stupka, E. & Rossiter, S.J. (2013). Genome-wide signatures of convergent evolution in echolocating mammals. Nature, 502, 228-231.

Parra, G., Bradnam, K. & Korf, I. (2007) CEGMA: a pipeline to accurately annotate core genes in eukaryotic genornes. Bioinformatics, 23(9), 1061-1067.

Paterson, A. H., Wendel, J. F., Gundlach, H., et al. (2012) Repeated polyploidization of *Gossypium* genomes and the evolution of spinnable cotton fibres. Nature, 492(7429), 423.

Patwardhan, A., Ray, S., Roy, A. 2014. Molecular Markers in Phylogenetic Studies-A Review. Journal of Phylogenetics and Evolutionary Biology, 2:2.

Pautasso, M., Aas, G., Queloz, V. & Holdenrieder, O. (2013) European ash (*Fraxinus excelsior*) dieback - A conservation biology challenge. Biological Conservation, 158, 37-49.

Pellicer, J., Fay, M. F. & Leitch, I. J. (2010) The largest eukaryotic genome of them all? Botanical Journal of the Linnean Society, 164(1), 10-15.

Peng, Q. & Ecker, J. R. (2012) Detection of allele-specific methylation through a generalized

heterogeneous epigenome model. Bioinformatics, 28(12), I163-I171.

Petit, R. J., Aguinagalde, I., de Beaulieu, J. L., et al. (2003) Glacial refugia: Hotspots but not melting pots of genetic diversity. Science, 300(5625), 1563-1565.

Petit, R. J., Brewer, S., Bordacs, S., et al. (2002) Identification of refugia and post-glacial colonisation routes of European white oaks based on chloroplast DNA and fossil pollen evidence. Forest Ecology and Management, 156(1-3), 49-74.

Petritan, A. M., Von Luepke, B. & Petritan, I. C. (2007) Effects of shade on growth and mortality of maple (*Acer pseudoplatanus*), ash (*Fraxinus excelsior*) and beech (*Fagus sylvatica*) saplings. Forestry, 80(4), 397-412.

Piotti, A., Leonarduzzi, C., Postolache, D., Bagnoli1, F., Spanu, I., Brousseau, L., Urbinati, C., Leonardi, S. & Vendramin, G.G. (2017). Unexpected scenarios from Mediterranean refugial areas: disentangling complex demographic dynamics along the Apennine distribution of silver fir. Journal of Biogeography, early view, DOI:10.1111/jbi.13011.

Platt, A., Gugger, P. F., Pellegrini, M. & Sork, V. L. (2015) Genome-wide signature of local adaptation linked to variable CpG methylation in oak populations. Molecular Ecology, 24(15), 3823-3830.

Pritchard, J. K., Stephens, M. & Donnelly, P. (2000) Inference of population structure using multilocus genotype data. Genetics, 155(2), 945-959.

Queloz, V., Grunig, C. R., Berndt, R., Kowalski, T., Sieber, T. N. & Holdenrieder, O. (2011) Cryptic speciation in *Hymenoscyphus albidus*. Forest Pathology, 41(2), 133-142.

Ramsay, J. & Schemske, D.W. (1998) PATHWAYS, MECHANISMS, AND RATES OF POLYPLOID FORMATION IN FLOWERING PLANTS. Annual Review of Ecology and Systematics, 29, 467-501.
Ranade, S.S., Garcia-Gil, M.R., Rossello, J.A. (2016). Non-functional plastid *ndh* gene fragments are present in the nuclear genome of Norway spruce (*Picea abies* L. Karsch): insights from in silico analysis of nuclear and organellar genomes. Molecular Genetics and Genomics, 291(2), 935-941.

Rebek, E. J., Herms, D. A. & Smitley, D. R. (2008) Interspecific variation in resistance to Emerald ash borer (Coleoptera : Buprestidae) among North American and Asian ash (*Fraxinus* spp.). Environmental Entomology, 37(1), 242-246.

Renny-Byfield, S., Gong, L., Gallagher, J. P. & Wendel, J. F. (2015) Persistence of Subgenomes in Paleopolyploid Cotton after 60 My of Evolution. Molecular Biology and Evolution, 32(4), 1063-1071.

Renny-Byfield, S. & Wendel, J. F. (2014) DOUBLING DOWN ON GENOMES: POLY-

PLOIDY AND CROP PLANTS. American Journal of Botany, 101(10), 1711-1725.

Rowley, E. R., Bryant, D. W., Fox, S. E., Givan, S. A., Mehlenbacher, S. A. & Mockler, T. C. (2014) Genome Sequencing and Resource Development for European Hazelnut. Viii International Congress on Hazelnut, 1052, 75-78.

Salmela, L. & Rivals, E. (2014) LoRDEC: accurate and efficient long read error correction. Bioinformatics, 30(24), 3506-3514.

Salmon, A., Ainouche, M. L. & Wendel, J. F. (2005) Genetic and epigenetic consequences of recent hybridization and polyploidy in *Spartina* (Poaceae). Molecular Ecology, 14(4), 1163-1175.

Schiffels, S. & Durbin, R. (2014) Inferring human population size and separation history from multiple genome sequences. Nature Genetics, 46(8), 919-925.

Schmitz, R. J., He, Y. P., Valdes-Lopez, O., Khan, S. M., Joshi, T., Urich, M. A., Nery, J. R., Diers, B., Xu, D., Stacey, G. & Ecker, J. R. (2013a) Epigenome-wide inheritance of cytosine methylation variants in a recombinant inbred population. Genome Research, 23(10), 1663-1674.

Schmitz, R. J., Schultz, M. D., Lewsey, M. G., O'Malley, R. C., Urich, M. A., Libiger, O., Schork, N. J. & Ecker, J. R. (2011) Transgenerational Epigenetic Instability Is a Source of Novel Methylation Variants. Science, 334(6054), 369-373.

Schmitz, R. J., Schultz, M. D., Urich, M. A., et al. (2013b) Patterns of population epigenomic diversity. Nature, 495(7440), 193-198.

Schmutz, J., Cannon, S. B., Schlueter, J., et al. (2010) Genome sequence of the palaeopolyploid soybean. Nature, 463(7278), 178-183.

Schnable, J. C., Pedersen, B. S., Subramaniam, S. & Freeling, M. (2011a) Dose-sensitivity, conserved non-coding sequences, and duplicate gene retention through multiple tetraploidies in the grasses. Frontiers in Plant Science, 2.

Schnable, J. C., Springer, N. M. & Freeling, M. (2011b) Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss. Proceedings of the National Academy of Sciences of the United States of America, 108(10), 4069-4074.

Schultz, M. D., Schmitz, R. J. & Ecker, J. R. (2012) 'Leveling' the playing field for analyses of single-base resolution DNA methylomes. Trends in Genetics, 28(12), 583-585.

Schwanda, K. & Kirisits, T. (2016) Pathogenicity of *Hymenoscyphus fraxineus* towards leaves of three European ash species: *Fraxinus excelsior*, *F. angustifolia* and *F. ornus*. Plant Pathology, 65(7), 1071-1083.

Sehrish, T., Symonds, V. V., Soltis, D. E., Soltis, P. S. & Tate, J. A. (2014) Gene silencing via DNA methylation in naturally occurring *Tragopogon miscellus* (Asteraceae) allopolyploids. BMC Genomics, 15.

Serapiglia, M. J., Gouker, F. E. & Smart, L. B. (2014) Early selection of novel triploid hybrids of shrub willow with improved biomass yield relative to diploids. BMC Plant Biology, 14.

Sha, A. H., Lin, X. H., Huang, J. B. & Zhang, D. P. (2005) Analysis of DNA methylation related to rice adult plant resistance to bacterial blight based on methylation-sensitive AFLP (MSAP) analysis. Molecular Genetics and Genomics, 273(6), 484-490.

Shaw, J., Lickey, E. B., Schilling, E. E. & Small, R. L. (2007) Comparison of whole chloroplast genome sequences to choose noncoding regions for phylogenetic studies in angiosperms: The tortoise and the hare III. American Journal of Botany, 94(3), 275-288.

Shen, Z. D., Sun, J., Yao, J., et al. (2015) High rates of virus- induced gene silencing by tobacco rattle virus in *Populus*. Tree Physiology, 35(9), 1016-1029.

Silva-Junior, O.B. & Grattapaglia, D. (2015) Genome-wide patterns of recombination, linkage disequilibrium and nucleotide diversity from pooled resequencing and single nucleotide polymorphism genotyping unlock the evolutionary history of *Eucalyptus grandis*. New Phytologist, 208(3), 830-834.

Simao, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. (2015) BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics, 31(19), 3210-3212.

Simpson, J. T., Wong, K., Jackman, S. D., Schein, J. E., Jones, S. J. M. & Birol, I. (2009) ABySS: A parallel assembler for short read sequence data. Genome Research, 19(6), 1117-1123.

Singh, R., Ong-Abdullah, M., Low, E. T. L., et al. (2013) Oil palm genome sequence reveals divergence of interfertile species in Old and New Worlds. Nature, 500(7462), 335.

Slavov, G.T., DiFazio, D.P. Martin, J., et al. (2012) Genome resequencing reveals multiscale geographic structure and extensive linkage disequilibrium in the forest tree *Populus trichocarpa*. New Phytologist, 196(3), 713-725.

Soltis, P. S., Marchant, D. B., Van de Peer, Y. & Soltis, D. E. (2015) Polyploidy and genome evolution in plants. Current Opinion in Genetics & Development, 35, 119-125.

Soltis, P. S. & Soltis, D. E. (2013) A conifer genome spruces up plant phylogenomics. Genome Biology, 14(6).

Song, K. M., Lu, P., Tang, K. L. & Osborn, T. C. (1995) RAPID GENOME CHANGE IN SYNTHETIC POLYPLOIDS OF BRASSICA AND ITS IMPLICATIONS FOR POLYPLOID EVOLUTION. Proceedings of the National Academy of Sciences of the United States of America, 92(17), 7719-7723.

Springer, N. M. (2013) Epigenetics and crop improvement. Trends in Genetics, 29(4), 241-247.

Su, C., Wang, C., He, L., Yang, C. P. & Wang, Y. C. (2014) Shotgun Bisulfite Sequencing of the *Betula platyphylla* Genome Reveals the Tree's DNA Methylation Patterning. International Journal of Molecular Sciences, 15(12), 22874-22886.

Sutherland, B. G., Belaj, A., Nier, S., Cottrell, J. E., Vaughan, S. P., Hubert, J. & Russell, K. (2010) Molecular biodiversity and population structure in common ash (*Fraxinus excelsior* L.) in Britain: implications for conservation. Molecular Ecology, 19(11), 2196-2211.

Suyama, M., Torrents, D. & Bork, P. (2006) PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Research, 34, W609-W612.

Thavamanikumar, S., McManus, L. J., Tibbits, J. F. G. & Bossinger, G. (2011) The significance of single nucleotide polymorphisms (SNPs) in *Eucalyptus globulus* breeding programs. Australian Forestry, 74(1), 23-29.

Thavamanikumar, S., Southerton, S.G., Bossinger, G., & Thumma, B.R. (2013). Dissection of complex traits in forest trees opportunities for marker-assisted selection. Tree Genetics & Genomes, 9(3), 627-639.

The Potato Genome Sequencing Consortium (2011) Genome sequence and analysis of the tuber crop potato. Nature, 475, 189-195.

The Tomato Genome Consortium (2012) The tomato genome sequence provides insights into fleshy fruit evolution. Nature, 465, 635-641.

Thomas, P. A. (2016) Biological Flora of the British Isles: *Fraxinus excelsior*. Journal of Ecology, 104(4), 1158-1209.

Thomasset, M., Fernandez-Manjarres, J. F., Douglas, G. C., Bertolino, P., Frascaria-Lacoste, N. & Hodkinson, T. R. (2013) Assignment testing reveals multiple introduced source populations including potential ash hybrids (*Fraxinus excelsior x F. angustifolia*) in Ireland. European Journal of Forest Research, 132(2), 195-209.

Thomasset, M., Fernandez-Manjarres, J. F., Douglas, G. C., Frascaria-Lacoste, N., Raquin, C. & Hodkinson, T. R. (2011) MOLECULAR AND MORPHOLOGICAL CHARACTERI-

ZATION OF RECIPROCAL F-1 HYBRID ASH (*FRAXINUS EXCELSIOR x FRAXINUS ANGUSTIFOLI*A, OLEACEAE) AND PARENTAL SPECIES REVEALS ASYMMETRIC CHARACTER INHERITANCE. International Journal of Plant Sciences, 172(3), 423-433.

Thompson, H. (2012) The chestnut resurrection. Nature, 490(7418), 22-23.

Timmermann, V., Brja, I., Hietala, A. M., Kirisits, T. & Solheim, H. (2011), Ash dieback: pathogen spread and diurnal patterns of ascospore dispersal, with special emphasis on Norway. EPPO Bulletin, 41: 1420.

Tollesfrud, M.M., Mykring, T., Sonstebo, J.H., Lygis, V., Hietala, A.M., & Heuertz, M. (2016). Genetic Structure in the Northern Range Margins of Common Ash, *Fraxinus excelsior* L. PLoS ONE, 11(12): e0167104.

Tulik, M., Marciszewska, K. & Adamczyk, J. (2010) Diminished vessel diameter as a possible factor in the decline of European ash (*Fraxinus excelsior* L.). Annals of Forest Science, 67(1).

Tuskan, G. A., DiFazio, S., Jansson, S., et al. (2006) The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). Science, 313(5793), 1596-1604.

Tuskan, G. A., DiFazio, S. P. & Teichmann, T. (2004) Poplar genomics is getting popular: The impact of the poplar genome project on tree research. Plant Biology, 6(1), 2-4.

USDA (2016). Webpage: https://www.aphis.usda.gov/aphis/ourfocus/planthealth/plant-pest-and-disease-programs/pests-and-diseases/emerald-ash-borer/ Last accessed on 21 November 2016.

VanBuren, R., Bryant, D., Edger, P. P., et al. (2015) Single-molecule sequencing of the desiccation-tolerant grass *Oropetium thomaeum*. Nature, 527(7579), 508.

Velasco, R., Zharkikh, A., Affourtit, J., et al. (2010) The genome of the domesticated apple (*Malus x domestica* Borkh.). Nature Genetics, 42(10), 833.

Verde, I., Abbott, A. G., Scalabrin, S., et al. (2013) The high-quality draft genome of peach (*Prunus persica*) identifies unique patterns of genetic diversity, domestication and genome evolution. Nature Genetics, 45(5), 487.

Verhoeven, K. J. F., Jansen, J. J., van Dijk, P. J. & Biere, A. (2010) Stress-induced DNA methylation changes and their heritability in asexual dandelions. New Phytologist, 185(4), 1108-1118.

Vezzi, F., Narzisi, G. & Mishra, B. (2012) Reevaluating Assembly Evaluations with Feature Response Curves: GAGE and Assemblathons. Plos One, 7(12).

Volk, T.A., Heavey, J.P., & Eisenbies, M.H. (2016). Advances in shrub-willow crops for

bioenergy, renewable products, and environmental benefits. Food and Energy Biosecurity, doi: 10.1002/fes3.82.

Wang, H. F., Beyene, G., Zhai, J. X., Feng, S. H., Fahlgren, N., Taylor, N. J., Bart, R., Carrington, J. C., Jacobsen, S. E. & Ausin, I. (2015a) CG gene body DNA methylation changes and evolution of duplicated genes in cassava. Proceedings of the National Academy of Sciences of the United States of America, 112(44), 13729-13734.

Wang, J., Street, N. R., Scofield, D. G. & Ingvarsson, P. K. (2016) Natural Selection and Recombination Rate Variation Shape Nucleotide Polymorphism Across the Genomes of Three Related Populus Species. Genetics, 202(3), 1185.

Wang, N., Thomson, M., Bodles, W. J. A., Crawford, R. M. M., Hunt, H. V., Featherstone, A. W., Pellicer, J. & Buggs, R. J. A. (2013) Genome sequence of dwarf birch (*Betula nana*) and cross-species RAD markers. Molecular Ecology, 22(11), 3098-3111.

Wang, P. F., Xia, H., Zhang, Y., Zhao, S. Z., Zhao, C. Z., Hou, L., Li, C. S., Li, A. Q., Ma, C. X. & Wang, X. J. (2015b) Genome-wide high-resolution mapping of DNA methylation identifies epigenetic variation across embryo and endosperm in Maize (*Zea may*). BMC Genomics, 16.

Warren, R. L., Keeling, C. I., Yuen, M. M. S., et al. (2015) Improved white spruce (*Picea glauca*) genome assemblies and annotation of large gene families of conifer terpenoid and phenolic defense metabolism. Plant Journal, 83(2), 189-212.

Wendel, J. F. (2015) The wondrous cycles of polyploidy in plants. American Journal of Botany, 102(11), 1753-1756.

Wendel, J. F., Jackson, S. A., Meyers, B. C. & Wing, R. A. (2016) Evolution of plant genome architecture. Genome Biology, 17.

Whitehill, J.G.A., Popova-Butler, A., Green-Church, K.B., Koch, J.L., Herms, D.A., Bonello, P. (2016) Interspecific Proteomic Comparisons Reveal Ash Phloem Genes Potentially Involved in Constitutive Resistance to the Emerald Ash Borer. PLoS ONE, 6(9): e24863.

Wilson, M. E., Maksaev, G. & Haswell, E. S. (2013) MscS-like Mechanosensitive Channels in Plants and Microbes. Biochemistry, 52(34), 5708-5722.

Wjkiewicza, B., & Wachowiaka, W. (2016) Substructuring of Scots pine in Europe based on polymorphism at chloroplast microsatellite loci. Flora - Morphology, Distribution, Functional Ecology of Plants, 220, 142-149.

Wu, G. A., Prochnik, S., Jenkins, J., et al. (2014) Sequencing of diverse mandarin, pummelo and orange genomes reveals complex history of admixture during citrus domestication. Nature Biotechnology, 32(7), 656.

Xu, Q., Chen, L. L., Ruan, X. A., et al. (2013) The draft genome of sweet orange (*Citrus sinensis*). Nature Genetics, 45(1), 59.

Zhang, H. M. & Zhu, J. K. (2011) RNA-directed DNA methylation. Current Opinion in Plant Biology, 14(2), 142-147.

Zhang, W. P. (2000) Phylogeny of the grass family (Poaceae) from rpl16 intron sequence data. Molecular Phylogenetics and Evolution, 15(1), 135-146.

Zhao, Y. J., Hosoya, T., Baral, H. O., Hosaka, K. & Kakishima, M. (2012) *Hymenoscyphus pseudoalbidus*, the correct name for *Lambertella albida* reported from Japan. Mycotaxon, 122, 25-41.

Zhong, S. L., Fei, Z. J., Chen, Y. R., et al. (2013) Single-base resolution methylomes of tomato fruit development reveal epigenome modifications associated with ripening. Nature Biotechnology, 31(2), 154-159.

Zimin, A., Stevens, K. A., Crepeau, M., et al. (2014) Sequencing and Assembly of the 22-Gb Loblolly Pine Genome. Genetics, 196(3), 875-890.

Zimin, A. V., Marcais, G., Puiu, D., Roberts, M., Salzberg, S. L. & Yorke, J. A. (2013) The MaSuRCA genome assembler. Bioinformatics, 29(21), 2669-2677.

Zohren, J., Wang, N. A., Kardailsky, I., Borrell, J. S., Joecker, A., Nichols, R. A. & Buggs, R. J. A. (2016) Unidirectional diploid-tetraploid introgression among British birch trees with shifting ranges shown by restriction site-associated markers. Molecular Ecology, 25(11), 2413-2426.

Appendix

1. Article on ash genome project, published in Nature on 12th January 2017 (Vol 541, 212-216), with online pre-print available from 26th December 2016.

2. Book chapter on emerging angiosperm tree genome projects, published by Springer as part of the Plant Genetics and Genomics: Crops and Models series. Available online from 31st December 2016.

LETTER

OPEN

doi:10.1038/nature20786

Genome sequence and genetic diversity of European ash trees

Elizabeth S. A. Sollars[1,2]*, Andrea L. Harper[3]*, Laura J. Kelly[1]*, Christine M. Sambles[4]*, Ricardo H. Ramirez-Gonzalez[5], David Swarbreck[5], Gemy Kaithakottil[5], Endymion D. Cooper[1], Cristobal Uauy[6], Lenka Havlickova[3], Gemma Worswick[1,8], David J. Studholme[4], Jasmin Zohren[1], Deborah L. Salmon[4], Bernardo J. Clavijo[5], Yi Li[3], Zhesi He[3], Alison Fellgett[3], Lea Vig McKinney[7], Lene Rostgaard Nielsen[7], Gerry C. Douglas[8], Erik Dahl Kjær[7], J. Allan Downie[6], David Boshier[9], Steve Lee[10], Jo Clark[11], Murray Grant[4]†, Ian Bancroft[3], Mario Caccamo[5,12] & Richard J. A. Buggs[1,13]

Ash trees (genus *Fraxinus*, family Oleaceae) are widespread throughout the Northern Hemisphere, but are being devastated in Europe by the fungus *Hymenoscyphus fraxineus*, causing ash dieback, and in North America by the herbivorous beetle *Agrilus planipennis*[1,2]. Here we sequence the genome of a low-heterozygosity *Fraxinus excelsior* tree from Gloucestershire, UK, annotating 38,852 protein-coding genes of which 25% appear ash specific when compared with the genomes of ten other plant species. Analyses of paralogous genes suggest a whole-genome duplication shared with olive (*Olea europaea*, Oleaceae). We also re-sequence 37 *F. excelsior* trees from Europe, finding evidence for apparent long-term decline in effective population size. Using our reference sequence, we re-analyse association transcriptomic data[3], yielding improved markers for reduced susceptibility to ash dieback. Surveys of these markers in British populations suggest that reduced susceptibility to ash dieback may be more widespread in Great Britain than in Denmark. We also present evidence that susceptibility of trees to *H. fraxineus* is associated with their iridoid glycoside levels. This rapid, integrated, multidisciplinary research response to an emerging health threat in a non-model organism opens the way for mitigation of the epidemic.

We sequenced a European ash (*F. excelsior*) tree generated from self-pollination of a woodland tree in Gloucestershire, UK. The sequenced tree (Earth Trust accession number 2451S) appeared free of ash dieback (ADB) when sampled in 2013 and 2014, but showed symptoms in February 2016. The haploid genome size was measured by flow cytometry as 877.24 ± 1.41 megabase pairs (Mbp). Total genomic DNA was sequenced to 192× coverage (see Supplementary Table 1). We assembled the genome into 89,514 nuclear scaffolds with an N_{50} (the length at which scaffolds include half the bases of the assembly) of 104 kilobase pairs (kbp), 26 mitochondrial scaffolds, and one plastid chromosome (Supplementary Tables 2 and 3), where the non-N assembly constitutes 80.5% of the predicted genome size. RepeatMasker estimated 35.90% of the assembly to be repetitive elements, with long terminal repeat retrotransposons predominating (Supplementary Table 4). Compared with other eudicot genomes of similar size[4,5] this repeat content is low. The 17% of the assembly composed of undetermined bases probably contains additional repeats; 27% of reads that do not map to the assembly align to ash repeats (Supplementary Table 5). We generated approximately 160 million RNA sequencing (RNA-seq) read pairs from tree 2451S leaf tissue and from leaf, cambium, root and flower tissue of its parent tree (Supplementary Table 6); low expression of repetitive elements was found in all tissues (Supplementary Table 7).

We annotated the genome using an evidence-based workflow incorporating protein and RNA-seq data, predicting 38,852 protein-coding genes and 50,743 transcripts (Supplementary Table 4). This gene count is within 12% that of tomato (version of genome (v)2.3)[4], potato (v3.4)[6] and hot pepper (v1.5)[7] but higher than monkey flower (v2.0; 26,718 genes)[8]. Evidence for completeness and coherence of our models is shown in Extended Data Fig. 1. Of 38,852 predicted genes, 97.67% (and 98.18% of transcripts) were supported by ash RNA-seq data, 81.80% showed high similarity to plant proteins (>50% high-scoring segment pair coverage) (Supplementary Table 8), 97.05% had matches in the non-redundant databases (excluding hits to ash), 82.74% generated hits to InterPro signatures and 78.09% were assigned Gene Ontology terms. We also identified 107 microRNA (miRNA), 792 transfer RNA (tRNA) and 51 ribosomal RNA (rRNA) genes.

Past whole-genome duplication events are commonly inferred from the distributions of pairwise synonymous site divergence (K_s) within paralogous gene groups[9]. We plotted these for ash and six other plant species (Fig. 1a and Supplementary Table 9). Ash and olive shared a peak near $K_s = 0.25$, suggesting an Oleaceae-specific whole-genome duplication. A peak near $K_s = 0.6$ shared by ash, olive, monkey flower and tomato but not by bladderwort, coffee and grape does not fit a common origin hypothesis, unless bladderwort has an accelerated substitution rate and the tomato peak is not restricted to the Solanales as evidenced previously[4]. Synteny analysis between ash and monkey flower did not provide conclusive evidence for shared whole-genome duplication (Extended Data Fig. 2). Duplicated genes in the ash genome that were not locally duplicated (that is, within ten genes of each other in our assembly) show no significantly enriched Gene Ontology terms at a false discovery rate level of 0.05. By contrast 1,005 locally duplicated genes showed significant enrichment of terms relating to oxidoreductase, catalytic and monooxygenase activity compared with all other genes, suggesting evolution of secondary metabolism by local duplications.

We analysed gene families shared between ash and 10 other species (Supplementary Table 10). In total, 279,603 proteins (77.14% of the input sequences) clustered into 27,222 groups, of which 4,292 contained sequences from all species, 3,266 were angiosperm-specific and 462 Eudicot-specific. Patterns of gene-family sharing among asterids and among woody species are shown in Fig. 1b, c. For 38,852 ash proteins,

[1]School of Biological and Chemical Sciences, Queen Mary University of London, Mile End Road, London E1 4NS, UK. [2]QIAGEN Aarhus A/S, Silkeborgvej 2, Prismet, 8000 Aarhus C., Denmark. [3]Centre for Novel Agricultural Products, University of York, Heslington, York YO10 5DD, UK. [4]Biosciences, College of Life and Environmental Sciences, University of Exeter, Exeter EX4 4QD, UK. [5]Earlham Institute, Norwich Research Park, Norwich NR4 7UZ, UK. [6]John Innes Centre, Norwich Research Park, Norwich NR4 7UH, UK. [7]Department of Geosciences and Natural Resource Management, University of Copenhagen, Rolighedsvej 23, 1958 Frederiksberg C, Denmark. [8]Teagasc, Agriculture and Food Development Authority, Ashtown, Dublin D15 KN3K, Ireland. [9]Department of Plant Sciences, University of Oxford, Oxford OX1 3RB, UK. [10]Forest Research, Northern Research Station, Roslin, Midlothian EH25 9SY, UK. [11]Earth Trust, Little Wittenham, Abingdon, Oxfordshire OX14 4QZ, UK. [12]National Institute of Agricultural Botany, Cambridge CB3 0LE, UK. [13]Royal Botanic Gardens Kew, Richmond, Surrey TW9 3AB, UK. †Present address: School of Life Sciences, Gibbet Hill Campus, University of Warwick, Coventry CV4 7AL, UK.
*These authors contributed equally to this work.

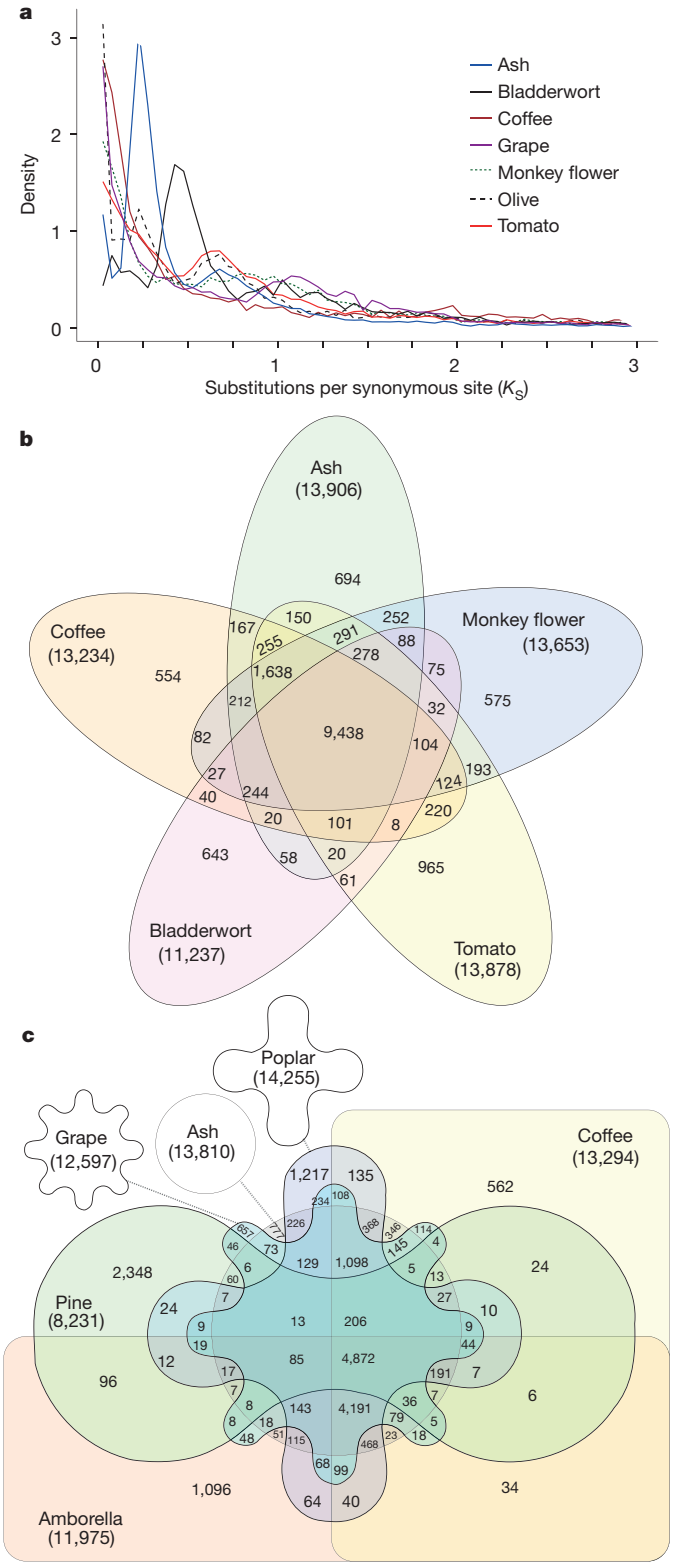


Figure 1 | Gene sharing within and among plant genomes. **a**, Distribution of K_s values between paralogous gene pairs within the genomes of ash (*F. excelsior*), tomato (*Solanum lycopersicum*), coffee (*Coffea canephora*), bladderwort (*Utricularia gibba*), grape (*Vitis vinifera*) and monkey flower (*Mimulus guttatus*), and transcriptome of olive (*O. europaea*). **b**, Venn diagram of gene sharing by five asterid species. **c**, Venn diagram of gene sharing by six woody species. Numbers in parentheses are the total number of OrthoMCL groups found for that species; numbers in intersections show the total number of groups shared between given combinations of taxa.

30,802 clustered into 14,099 groups, of which 643 were ash-specific, containing 1,554 proteins. There were also 8,050 singleton proteins unique to ash. Of the 9,604 ash-specific proteins, 6,405 matched at least one InterPro signature. The 20 largest groups in ash are listed in Extended Data Table 1: several are putatively associated with disease resistance.

To investigate genomic diversity in *F. excelsior*, we sequenced 37 ash trees from central, northern and western Europe (Fig. 2 and Supplementary Table 11), to an average of 8.4× genome coverage by trimmed and filtered reads. Together with reads from Danish 'Tree35' (http://oadb.tsl.ac.uk/), these were mapped to the reference genome. We found 12.48 million polymorphic sites with a variant of high confidence in at least one individual (quality > 300 using freebayes[10]): we refer to these as the 'genome-wide SNP set' in the 'European Diversity Panel'. Of these, 6.85 million (54.88%) occur inside or within 5 kbp of genes (Supplementary Table 12). We found 259,946 amino-acid substitutions and 71,513 variants that affect stop or start codons, or splice sites. We selected 23 amino-acid variants, and 26 non-coding variants from the 'genome-wide SNP set' with a range of call qualities for validation using KASP: individual genotype calls with quality greater than 300 have a false-positive rate of 6% and those with quality greater than 1,000 have a false-positive rate of zero (Supplementary Table 13). We ran a more stringent variant calling restricted to regions of the genome with between 5× and 30× coverage in all 38 samples. These totalled 20.6 Mbp (2.3% of the genome), within which 529,812 variants were called with CLC Genomics Workbench. Of these, 394,885 were bi-allelic single nucleotide polymorphisms (SNPs) with minimum allele frequency above 0.05, which we refer to as the 'reduced SNP set'. We also found about 31,300 singleton simple sequence repeat (SSR) loci in the ash genome, and designed primers for 664 (Supplementary Data 1). In a sample of 366 of these, 48% were polymorphic in the European Diversity Panel sequences. We PCR tested 48 of these in multiplexes with European Diversity Panel genomic DNA and found that 41 amplified successfully (Supplementary Data 1).

We analysed population structure of the European Diversity Panel using a plastid haplotype network; STRUCTURE[11] runs on genomic SNPs and principal component analysis (PCA) of the 'reduced SNP set' (Fig. 2a–d and Extended Data Fig. 3). Clearest differentiation was found in the plastid network, with four distinct haplotype groups each separated from each other by at least 20 substitutions. One group was more frequent in Great Britain than on the continental Europe. The second and third principal components of the PCA corresponded with the plastid data somewhat (Fig. 2c). Previous analyses of SSRs in plastids identified variants unique to the British Isles and Iberia[12]. Linkage disequilibrium in the European Diversity Panel decayed logarithmically, with an average r^2 of 0.15 at 100 bp between SNPs, reaching an r^2 of 0.05 at ~40 kbp (Fig. 2e). This is similar to long-range linkage disequilibrium estimates found in *Populus tremuloides*[13]. An apparent long-term effective population size decline of *F. excelsior* in Europe was shown by analyses based on heterozygosity in the reference genome (using pairwise sequentially Markovian coalescent (PSMC)[14], Fig. 2f). Such patterns may also reflect a complex history of population subdivision in ash[15].

We used associative transcriptomics to predict ADB damage in Great Britain. We used the full coding DNA sequence (CDS) models from our genome annotation as a mapping reference for previously generated[3] RNA-seq reads from 182 Danish ash accessions ('Danish Scored Panel') that have been exposed to *H. fraxineus*, and scored for damage (Supplementary Data 2). This yielded 40,133 gene expression markers (GEMs; Supplementary Data 3) and 394,006 SNPs (Supplementary Data 4). Twenty GEMs were associated with ADB damage scores, including eight MADS-box proteins, and two cinnamoyl-CoA reductase 2 genes that may be involved in the hypersensitive response (Supplementary Data 5). Four assays representing the top five GEMs were applied to 58 Danish accessions ('Danish Test Panel') to validate the top markers. Results were combined into a single predicted damage

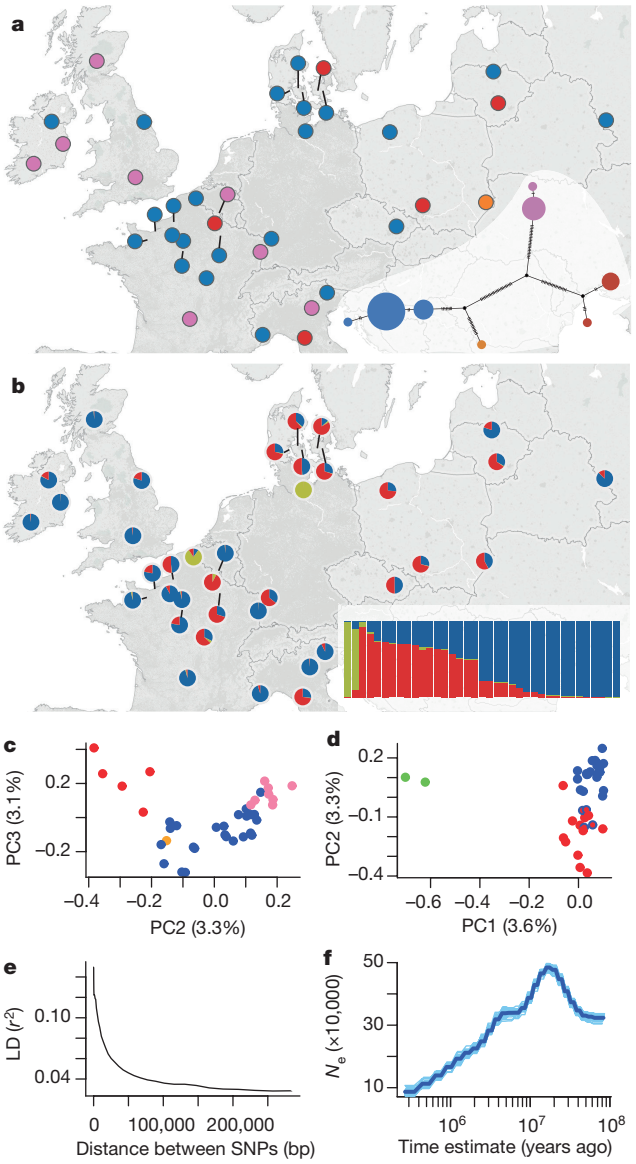


Figure 2 | Genome diversity of *F. excelsior* in Europe. a, Map showing the distribution of plastid haplotypes ($n = 37$), on the basis of a median-joining plastid haplotype network for the European Diversity Panel (inset). **b**, Map showing diversity structure of genomic SNPs, on the basis of average *Q* value for each individual (inset), from three runs of STRUCTURE with different sets of 8,955 SNPs and $k = 3$. **c**, PCA of 34,607 nuclear SNPs in the European Diversity Panel, PC2 plotted against PC3, with points coloured by plastid haplotype. **d**, From the same PCA, PC1 plotted against PC2, with points coloured by groupings found by STRUCTURE using genomic SNPs. **e**, Linkage disequilibrium decay between SNPs in the European Diversity Panel. **f**, Effective population size (n_e) history estimated using the PSMC method on the reference genome, with 100 bootstraps (shown in light blue).

score for each tree (Supplementary Data 6), which was compared with the observed damage scores (Fig. 3; $r^2 = 0.25$, $P = 6.9 \times 10^{-5}$): predictions of damage less than 50% consistently detected trees with very low observed damage scores. The same assays were also applied to 130 accessions from across the British range of *F. excelsior* ('British Screening Panel'; Supplementary Data 6). Strikingly, this provided lower predictions for ADB damage in the British Screening Panel: 25% were predicted to have <25% canopy damage compared with 9% of the Danish Test Panel. Trees with low predicted damage are scattered throughout Britain (Fig. 3).

We also examined expression of the top five GEM loci using reads per kilobase pair per million aligned reads (RPKM) values from our

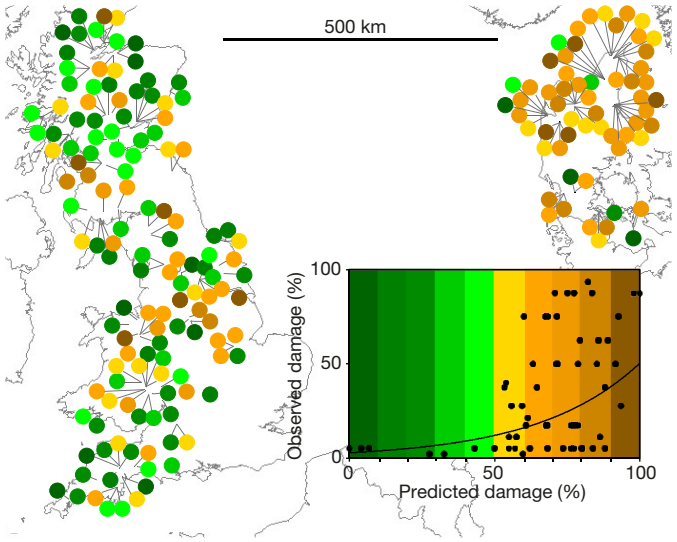


Figure 3 | Predicted ADB damage scores in Great Britain and Denmark. Map points are scaled by hue (high predicted damage scores in brown, low in green) and plotted according to the geographical origin of the parent trees of the British Screening Panel ($n = 130$) and the Danish Test Panel ($n = 58$). Single leaf samples taken from grafts of each individual tree were used for predicting damage scores. Inset: damage predictions for the Danish Test Panel ($n = 58$) correlated with log(mean observed damage scores) from 2013 to 2014 ($r^2 = 0.25$, $P = 6.9 \times 10^{-5}$).

shotgun Illumina read data for the reference tree (Extended Data Fig. 4), comparing these with RPKM values from the Danish Scoring Panel. Expression patterns in the reference tree were highly correlated with those of the most susceptible Danish quartile ($r^2 = 0.995$, $P < 0.001$), but not the least susceptible ($P = 0.24$), consistent with observations that the reference tree is now succumbing to the disease. We correlated the expression of all 20 top GEM markers in leaf, flower, cambium and root transcriptomes of the parent of the reference tree. This revealed that leaf expression levels were positively correlated with those in the cambium ($r^2 = 0.65$, $P < 0.001$) and flower ($r^2 = 0.38$, $P = 0.0041$), but not with the root ($P = 0.3594$).

We identified putative orthologues of the five GEM loci using our OrthoMCL results (Supplementary Data 5) and BLAST searches of GenBank, and conducted maximum likelihood and Bayesian analyses of relevant hits (Extended Data Fig. 5). FRAEX38873_v2_000173540.4, FRAEX38873_v2_000048340.1 and FRAEX38873_v2_000048360.1 clustered into the SVP/StMADS11 group[16] of type II MADS-box genes. FRAEX38873_v2_000261470.1 and FRAEX38873_v2_000199610.1 clustered into the SOC1/TM3 group of type II MADS-box proteins[16,17]. Both groups have roles in flower development[18–21], and appear to be involved in stress response in *Brassica rapa*[22]. Many genes involved in regulation of flowering time in *Arabidopsis thaliana* are involved in controlling phenology in perennial trees species[23], and genes belonging to the SVP/StMADS11 clade have potential roles in growth cessation, bud set and dormancy[23]. In *A. thaliana*, *AGL22*/*SVP* may be required for age-related resistance[24].

One mechanism by which transcriptional cascades, such as those involving MADS box genes, might be involved in tolerance or resistance to pathogens is via modulation of secondary metabolite concentrations. For five high-susceptibility and five low-susceptibility Danish trees, we profiled methanol-extracted leaf samples by liquid chromatography/mass spectrometry on a quadrupole time-of-flight mass spectrometer. Partial least squares discriminant analysis (PLS-DA) clearly discriminated high- and low-susceptibility trees (Fig. 4a). By using accurate mass to identify the chemical nature of discriminant features, we found greater abundance (Fig. 4b) of iridoid glycosides (for details see Extended Data Figs 6–9 and Supplementary Data 9) in genotypes with high susceptibility to ADB than in low-susceptibility genotypes.

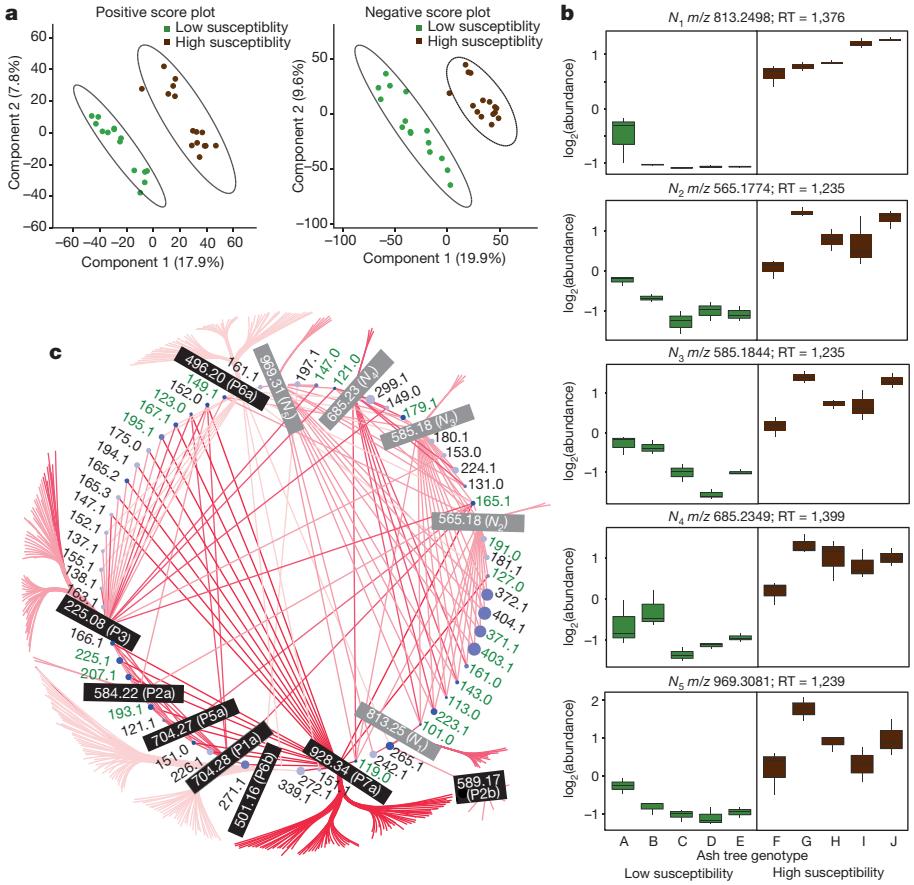


Figure 4 | Putative iridoid glycosides as discriminatory features between *F. excelsior* genotypes with differential susceptibility to ADB. **a**, Multivariate analysis PLS-DA score plot of metabolic profiles of five high-susceptibility and five low-susceptibility trees ($n = 3$ per genotype). **b**, Box-plots from these profiles showing normalized (internal standard) intensity ($\log_2$ transformed) of five discriminatory features observed in negative mode; m/z and retention time (RT) are given for each feature. **c**, Fragmentation network of discriminatory features, highlighted in black (positive mode) and grey (negative mode). Each product ion is labelled with its size (m/z), also depicted by its circle size. Blue shading increases with the number of times each ion is present in the precursor discriminatory features. Product ions not shared among precursors are shown as unlabelled tips. The edges are in shades of red on the basis of retention time; the paler the colour the earlier the retention time. Those fragment masses shaded in green have been previously reported from fragmentation of iridoid glycosides.

A tandem mass spectrometry (MS/MS) fragmentation network identified several product ions expected from fragmentation of iridoid glycosides (Fig. 4c). Iridoid glycosides are a well-known anti-herbivore defence mechanism in the Oleaceae[25–27]. They can also enhance fungal growth *in vitro*[28], although their aglycone hydrolysis product formed following tissue damage can also mediate fungal resistance[29]. Our data suggest there may be a trade-off between ADB susceptibility and herbivore susceptibility. This is of particular concern given the threat of *A. planipennis* to ash in both North America[1] and Europe[30] and may hamper efforts to breed trees with low susceptibility to both threats.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 23 February; accepted 11 November 2016.
Published online 26 December 2016.

1. Poland, T. M. & McCullough, D. G. Emerald ash borer: invasion of the urban forest and the threat to North America's ash resource. *J. Forest.* **104,** 118–124 (2006).
2. Pautasso, M., Aas, G., Queloz, V. & Holdenrieder, O. European ash (*Fraxinus excelsior*) dieback—a conservation biology challenge. *Biol. Conserv.* **158,** 37–49 (2013).
3. Harper, A. L. *et al.* Molecular markers for tolerance of European ash (*Fraxinus excelsior*) to dieback disease identified using associative transcriptomics. *Sci. Rep.* **6,** 19335 (2016).
4. The Tomato Genome Consortium. The tomato genome sequence provides insights into fleshy fruit evolution. *Nature* **485,** 635–641 (2012).
5. Ming, R. *et al.* Genome of the long-living sacred lotus (*Nelumbo nucifera* Gaertn.). *Genome Biol.* **14,** R41 (2013).
6. Potato Genome Sequencing Consortium. Genome sequence and analysis of the tuber crop potato. *Nature* **475,** 189–195 (2011).
7. Kim, S. *et al.* Genome sequence of the hot pepper provides insights into the evolution of pungency in *Capsicum* species. *Nature Genet.* **46,** 270–278 (2014).
8. Hellsten, U. *et al.* Fine-scale variation in meiotic recombination in *Mimulus* inferred from population shotgun sequencing. *Proc. Natl Acad. Sci. USA* **110,** 19478–19482 (2013).
9. Blanc, G. & Wolfe, K. H. Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. *Plant Cell* **16,** 1667–1678 (2004).
10. Garrison, E. & Marth, G. Haplotype-based variant detection from short-read sequencing. Preprint at https://arxiv.org/abs/1207.3907v2 [q-bio.GN] (2012).
11. Pritchard, J. K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics* **155,** 945–959 (2000).
12. Heuertz, M. *et al.* Chloroplast DNA phylogeography of European ashes, *Fraxinus* sp. (Oleaceae): roles of hybridization and life history traits. *Mol. Ecol.* **15,** 2131–2140 (2006).
13. Wang, J., Street, N. R., Scofield, D. G. & Ingvarsson, P. K. Natural selection and recombination rate variation shape nucleotide polymorphism across the genomes of three related *Populus* species. *Genetics* **202,** 1185–1200 (2016).
14. Li, H. & Durbin, R. Inference of human population history from individual whole-genome sequences. *Nature* **475,** 493–496 (2011).
15. Mazet, O., Rodríguez, W., Grusea, S., Boitard, S. & Chikhi, L. On the importance of being structured: instantaneous coalescence rates and human evolution—lessons for ancestral population size inference? *Heredity* **116,** 362–371 (2016).
16. Smaczniak, C., Immink, R. G. H., Angenent, G. C. & Kaufmann, K. Developmental and evolutionary diversity of plant MADS-domain factors: insights from recent studies. *Development* **139,** 3081–3098 (2012).

17. Wells, C. E., Vendramin, E., Jimenez Tarodo, S., Verde, I. & Bielenberg, D. G. A genome-wide analysis of MADS-box genes in peach [*Prunus persica* (L.) Batsch]. *BMC Plant Biol.* **15,** 41 (2015).
18. Liu, C. *et al.* Direct interaction of AGL24 and SOC1 integrates flowering signals in *Arabidopsis*. *Development* **135,** 1481–1491 (2008).
19. Li, D. *et al.* A repressor complex governs the integration of flowering signals in *Arabidopsis*. *Dev. Cell* **15,** 110–120 (2008).
20. Dorca-Fornell, C. *et al.* The *Arabidopsis* SOC1-like genes *AGL42*, *AGL71* and *AGL72* promote flowering in the shoot apical and axillary meristems. *Plant J.* **67,** 1006–1017 (2011).
21. Gregis, V., Sessa, A., Colombo, L. & Kater, M. M. *AGAMOUS-LIKE24* and SHORT *VEGETATIVE PHASE* determine floral meristem identity in Arabidopsis. *Plant J.* **56,** 891–902 (2008).
22. Saha, G. *et al.* Genome-wide identification and characterization of MADS-box family genes related to organ development and stress resistance in *Brassica rapa*. *BMC Genomics* **16,** 178 (2015).
23. Ding, J. & Nilsson, O. Molecular regulation of phenology in trees—because the seasons they are a-changin'. *Curr. Opin. Plant Biol.* **29,** 73–79 (2016).
24. Wilson, D. C., Carella, P., Isaacs, M. & Cameron, R. K. The floral transition is not the developmental switch that confers competence for the *Arabidopsis* age-related resistance response to *Pseudomonas syringae* pv. *tomato*. *Plant Mol. Biol.* **83,** 235–246 (2013).
25. Jensen, S. R., Franzyk, H. & Wallander, E. Chemotaxonomy of the Oleaceae: iridoids as taxonomic markers. *Phytochemistry* **60,** 213–231 (2002).
26. Kubo, I., Matsumoto, A. & Takase, I. A multichemical defense mechanism of bitter olive *Olea europaea* (Oleaceae): is oleuropein a phytoalexin precursor? *J. Chem. Ecol.* **11,** 251–263 (1985).
27. Eyles, A. *et al.* Comparative phloem chemistry of Manchurian (*Fraxinus mandshurica*) and two North American ash species (*Fraxinus americana* and *Fraxinus pennsylvanica*). *J. Chem. Ecol.* **33,** 1430–1448 (2007).
28. Marak, H. B., Biere, A. & Van Damme, J. M. M. Systemic, genotype-specific induction of two herbivore-deterrent iridoid glycosides in *Plantago lanceolata* L. in response to fungal infection by *Diaporthe adunca* (Rob.) Niessel. *J. Chem. Ecol.* **28,** 2429–2448 (2002).
29. Biere, A., Marak, H. B. & van Damme, J. M. M. Plant chemical defense against herbivores and pathogens: generalized defense or trade-offs? *Oecologia* **140,** 430–441 (2004).
30. Valenta, V., Moser, D., Kuttner, M., Peterseil, J. & Essl, F. A high-resolution map of emerald ash borer invasion risk for southern central Europe. *For. Trees Livelihoods* **6,** 3075–3086 (2015).

Supplementary Information is available in the online version of the paper.

Acknowledgements Eurofins MWG provided a discounted service for Illumina and 454 sequencing of the reference genome, funded by Natural Environment Research Council (NERC) Urgency Grant NE/K01112X/1 to R.J.A.B. The associative transcriptomic and metabolomic work was part of the 'Nornex' project led by J.A.D. funded jointly by the UK Biotechnology and Biological Sciences Research Council (BBSRC) (BBS/E/J/000CA5323) and the Department for Environment, Food & Rural Affairs. The Earlham Institute, Norwich, UK, sequenced 'Tree 35' funded by 'Nornex' and the European Diversity Panel funded by the Earlham Institute National Capability in Genomics (BB/J010375/1) grant. W. Crowther assisted with DNA extractions for the KASP assay; The John Innes Centre contributed KASP analyses. J. F. Miranda assisted with RNA extractions and quantitative PCR with reverse transcription (qRT–PCR) at the University of York. H. V. Florance, N. Smirnoff and the Exeter Metabolomics Facility developed metabolomic methods and ran samples, and T. P. Howard helped with statistics. L.J.K. and R.J.A.B. were partly funded by Living with Environmental Change (LWEC) Tree Health and Plant Biosecurity Initiative - Phase 2 grant BB/L012162/1 to R.J.A.B., S.L. and P. Jepson funded jointly by a grant from the BBSRC, Defra, Economic and Social Research Council, the Forestry Commission, NERC and the Scottish Government, under the Tree Health and Plant Biosecurity Initiative. G.W. was funded by Teagasc Walsh Fellowship 2014001 to R.J.A.B. and G.C.D. E.D.C. was funded by a Marie Skłodowska-Curie Individual Fellowship 'FraxiFam' (grant agreement 660003) to E.D.C. and R.J.A.B. E.S.A.S. and J.Z. were funded by the Marie Skłodowska-Curie Initial Training Network INTERCROSSING. J.A.D. received a John Innes Foundation fellowship. We thank A. Joecker for supervising E.S.A.S. at Qiagen and for helpful discussions. R.H.R.G. is supported by a Norwich Research Park PhD Studentship and Earlham Institute Funding and Maintenance Grant. This research used Queen Mary's MidPlus computational facilities, supported by QMUL Research-IT and funded by Engineering and Physical Sciences Research Council grant EP/K000128/1 and NERC EOS Cloud. D.J.S. acknowledges the support of BBSRC grant BB/N021452/1, which partly supported M.G., C.M.S. and D.J.S. during this work.

Author Contributions R.J.A.B., M.C., D.S., M.G., J.A.D. and I.B. are the lead investigators. R.J.A.B. coordinated the project and directed work on the reference genome. E.S.A.S. assembled the reference genome and organellar genomes, and analysed gene and genome duplications, European population structure and past effective population sizes. L.J.K. extracted high molecular mass DNA for the European Diversity Panel and conducted repetitive element, OrthoMCL and phylogenetic analyses. G.W. conducted SSR analyses. J.Z. extracted high molecular mass DNA and RNA for the reference genome. E.D.C. analysed genome duplication in the reference genome. D.S. and G.K. performed bioinformatic analyses to annotate the reference genome. M.C. conceived and, with R.J.A.B., oversaw the European Diversity Panel sequencing. R.R.-G., E.S.A.S. and M.C. performed SNP calling on the European Diversity Panel, and KASP genotyping. C.U. conducted KASP genotyping. B.J.C. conceived and oversaw the NEXTERA sequencing on the reference tree genome. M.C., J.A.D. and B.J.C. generated the first-pass 'Tree 35' Illumina reads included in the European-wide SNP analysis. E.D.K., L.R.N. and L.V.M. generated, selected and collected Danish samples. D.B. generated and J.C. maintained and sampled the reference tree. J.C., D.B., G.C.D. and S.L. generated, selected and collected UK and European Diversity Panel samples. For the associative transcriptomics, I.B. and A.L.H. conceived and planned the study; A.L.H., L.H. and A.F. performed experiments; bioinformatics was executed by Y.L. and Z.H.; and A.L.H. completed the data analysis. For the metabolomics, C.M.S., D.J.S. and M.G. conceived and conducted the analyses; C.M.S. developed methodology; and D.L.S. processed and extracted samples and ran the mass spectrometer.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to R.J.A.B. (r.buggs@qmul.ac.uk).

Reviewer Information *Nature* thanks P. Dorrestein, S. Jansson and the other anonymous reviewer(s) for their contribution to the peer review of this work.

METHODS

No statistical methods were used to predetermine sample size. The experiments were not randomized. The investigators were not blinded to allocation during experiments and outcome assessment.

Tree material. Reference tree. In 2013 twig material was collected from tree 2451S growing at Paradise Wood, Earth Trust, Oxfordshire, UK. This tree was produced via self-pollination of a hermaphroditic *F. excelsior* tree growing in woodland in Gloucestershire (latitude 52.020592, longitude −1.832804), UK, in 2002 as part of the FRAXIGEN project[31]. The parent tree was one of 19 trees that produced seed from self-pollination, and had lower heterozygosity at four microsatellite loci than the other 18 trees (D.B., unpublished observations). DNA was extracted from bud, cambial and wood tissues using CTAB[32] and Qiagen DNeasy protocols. RNA was extracted using the Qiagen RNeasy protocol from leaf tissue of tree 2451S and from leaf, cambium, root, and flower tissue of its parent tree in Gloucestershire.

European Diversity Panel. In 2014, twig material was collected from 37 trees representing 37 European provenances in a trial of *F. excelsior* established in 2004 at Paradise Wood, Earth Trust, Oxfordshire, UK, as part of the Realizing Ash's Potential project. DNA was extracted from cambial tissue of the twigs using a CTAB protocol.

British Screening Panel. In 2015, freshly flushed leaf material was collected from a clonal seed orchard of *F. excelsior* growing at Paradise Wood, Earth Trust, Oxfordshire, UK, for RNA extraction and complementary DNA (cDNA) synthesis as in ref. 3. Single whole leaves were harvested from four ramets of each of 130 ash trees selected from phenotypically superior parents throughout Britain, which had been cloned by grafting.

2451S DNA sequencing and genome assembly. The genome size of 2451S was estimated by flow cytometry with propidium iodide staining of nuclei, using leaf tissue co-chopped with an internal standard using a razor blade. Three preparations were made: two with *Petroselinum crispum* 'Curled Moss' parsley as standard (2C genome size = 4.50 pg)[33] and one with *S. lycopersicum* 'Stupicke polnı rane' (2C = 1.96 pg)[34] as standard. The Partec CyStain Absolut P protocol was used (Partec, Germany). Each preparation was measured six times, with the relative fluorescence of over 5,000 particles per replicate recorded on a Partec Cyflow SL3 (Partec, Germany) flow cytometer fitted with a 100-mW green solid state laser (Cobolt Samba; Cobolt, Sweden). The resulting histograms were analysed with the Flow-Max software (version 2.4, Partec). The measurement with the tomato internal standard was used as the best estimate of genome size, because the tomato genome size is closest to that of 2451S, yielding a more accurate result.

Genomic DNA of 2451S was sequenced using the following methods: (1) HiSeq 2000 (Illumina, San Diego, California, USA) at Eurofins, Ebersberg, Germany, with 100 bp reads and shotgun libraries with fragment sizes of 200 bp, 300 bp and 500 bp, and long jumping distance libraries with 3 kbp, 8 kbp, 20 kbp and 40 kbp insert sizes, generating 188× genome coverage; (2) 454 FLX+ (Roche, Switzerland) at Eurofins with shotgun libraries and maximum read length of 1,763 bp and mean length of 642 bp giving 4.3× genome coverage; and (3) MiSeq (Illumina, San Diego, California) at the Earlham Institute, Norwich, UK, with 300 bp paired-end reads from a Nextera library with ~5 kbp insert size, giving 16× genome coverage (see Supplementary Table 1). We assembled and released five genome assembly versions over the course of 3 years, details of which can be found in Supplementary Table 3. The most recent version assembled first into 235,463 contigs with a total size of 663 Mbp and an N_{50} of 5.7 kbp (Supplementary Table 2), and after scaffolding and removing organellar scaffolds, the assembly comprised 89,487 scaffolds totalling 867 Mbp (17% 'N') with an N_{50} of 104 kbp (Supplementary Table 2). The plastid genome was assembled separately into one circular contig of 155,498 bp, including an inverted repeat region of approximately 25,700 bp. The mitochondrial genome initially assembled into 296 contigs totalling 232 kbp. After several rounds of contig extension using overlaps of mapped 454 reads, the final assembly consisted of 26 contigs totalling 581 kbp with an N_{50} of 60.6 kbp.

All Illumina reads from 2451S were trimmed using CLC Genomics Workbench (QIAGEN Aarhus, Denmark) versions 6–8 (depending on when the data were received) to a minimum quality score of 0.01 (equivalent to Phred quality score of 20), a minimum length of 50 bp, and were trimmed of any adaptor and repetitive telomere sequences. The MiSeq Nextera reads were also run through FLASH[35] to merge overlapping paired reads, and NextClip[36] to remove adaptor sequences, both used with default parameters. Roche 454 reads were trimmed to a minimum Phred score of 0.05, and minimum length of 50 bp. *De novo* assembly was performed with the CLC Genomics Workbench, using the 200 bp, 300 bp, 500 bp and 5 kbp insert size Illumina library reads to build the De Bruijn graphs. The remaining Illumina reads and the 454 reads were used as 'guidance only reads' to help select the most supported path through the De Bruijn graphs. A word size (*k*-mer, a substring of length *k* in DNA sequence data) of 50 and maximum bubble size of 5,000 were used to assemble the reads into contigs with a minimum length of 500 bp. Contigs were then scaffolded with the stand-alone tool SSPACE[37] Basic version 2.0 using all paired Illumina reads, with the '-k' parameter (number of mapped paired reads required to join contigs) set to 7. Gaps in the scaffolds were closed using the GapCloser version 1.12 program using all paired reads (except for long jumping distance libraries), with pair_num_cutoff parameter set at 7. Four hundred and fifty-four reads were mapped to the assembly and used to join overlapping scaffolds using the Jelly.py script from PBSuite[38] version 14.7.14 with the following blasr parameters: -minMatch 11 -minPctIdentity 70 -bestn 1 -nCandidates 10 -maxScore -500 -noSplitSubreads. Contig57544 was removed from the assembly because it aligned fully to the PhiX bacteriophage genome, indicating it derived from the PhiX control library added to Illumina sequencing runs.

To assemble the plastid and mitochondrial genomes, high read depth 50 bp *k*-mers were extracted from the 200, 300 and 500 bp read libraries. Jellyfish[39] version 2.1.1 was used to count the depth for each *k*-mer, and these values were plotted in a scatterplot to identify peaks that could correspond to the organellar genomes. Every *k*-mer over 600× coverage was used in a BLAST search against the NCBI non-redundant (nr) database with a filter allowing only plant sequences; *k*-mers were then extracted on the basis of whether their first hit contained a 'mitochondrion' or 'plastid/chloroplast' related description. Reads from the 200, 300 and 500 bp libraries were then filtered against the *k*-mer sets, and were kept if the first and last 50 bp matched *k*-mers from the extracted sets (reads were at most 90 bp long). Each set of reads (mitochondrial and plastid) were then assembled *de novo* using the CLC Genomics Workbench. The plastid genome assembled initially into two contigs, which were joined using an alignment to the *O. europaea* plastid genome (GenBank accession number NC_015401.1), with the inverted repeat region being identified also. Reads from the 454 library were mapped to the assembly to check the sequence and especially the join region. The mitochondrial genome assembled first into 296 contigs. To fill in gaps and join the contigs together, 454 reads were mapped against the assembly and contig ends were extended using the Extend Contigs tool in the CLC Genome Finishing Module. The Join Contigs tool was then used to join overlapping ends together, and 454 reads were mapped to the resulting assembly to check any joined regions. Using this method of 'Map-Extend-Join' iteratively (approximately ten times in total), a more contiguous assembly of 26 contigs was obtained.

RNA sequencing. The five RNA samples (see 'Tree Material' above) were sequenced paired-end on Illumina HiSeq 2000 with 200 bp insert sizes, and a read length of 100 bp, at the QMUL Genome Centre, London, UK. Reads were trimmed using CLC Genomics Workbench to a minimum quality score of 0.01 (equivalent to Phred score of 20) and minimum length of 50 bp, and adaptors were also removed (Supplementary Table 6).

Analysis of repetitive DNA. The repetitive element (transposable elements and tandem repeats) content of the ash genome was analysed via two approaches: (1) *de novo* identification of the most abundant repeat families from unassembled 454 and Illumina reads; (2) *de novo* and similarity-based identification of repeats from the ash genome assembly.

***De novo* identification of repeat families from unassembled reads.** Individual 454 reads and Illumina read pairs from the 500 bp insert library (after adaptor trimming, but before any further quality control or filtering: see above) were used for *de novo* repeat identification. Reads were quality filtered and trimmed using the FASTX-Toolkit version 0.0.13 (http://hannonlab.cshl.edu/fastx_toolkit/index.html). Using fastx_trimmer, the first 10 bp of all reads (454 and Illumina) were removed (owing to skewed base composition). The 454 reads were clipped to a maximum of 250 bp and Illumina reads to a maximum of 90 bp; all shorter reads were removed using a custom Perl script. Reads were then quality filtered with the fastq_quality_filter tool to retain only those where 90% of bases had a Phred score of at least 20. Exact duplicates (which are probably artefacts from the emulsion PCR[40]) were removed from the 454 reads using the fastx_collapser tool.

The complete set of quality filtered and trimmed 454 reads (3,330,483) was used as input for the RepeatExplorer pipeline on Galaxy[41], with a minimum of 138 bp overlap for clustering and a minimum of 100 bp overlap for assembly. All clusters containing at least 0.01% of the input reads were examined manually to identify clusters that required merging (that is, where there was evidence that a single repeat family had been split over multiple clusters). Clusters were merged if they met the following three criteria: (1) they shared a significant number of similarity hits (for example, in a pair of clusters, 10% of the reads in the smaller cluster had BLAST hits to reads in the larger cluster); (2) they were the same repeat type (for example, LINEs); (3) they could be merged in a logical position (for example, for repetitive elements containing conserved domains these domains would be joined in the correct order). The re-clustering pipeline was run with a minimum of 100 bp overlap for assembly; merged clusters were examined manually to verify that all domains were in the correct orientation.

Quality filtered and trimmed Illumina reads were paired using the FASTA interlacer tool (version 1.0.0) in RepeatExplorer, resulting in 111,230,011 pairs;

unpaired reads were discarded. An initial run of RepeatExplorer with a sample of 100,000 read pairs was performed to obtain an estimate of the maximum number of reads that could be handled by the pipeline. A random sample of 3.5 million read pairs was then taken using the sequence sampling tool (version 1.0.0) in RepeatExplorer and used as input for the clustering pipeline, which further randomly subsampled the reads down to 3,370,186 pairs. The pipeline was run with a minimum of 50 bp overlap for clustering and a minimum of 36 bp overlap for assembly. Clusters containing at least 0.01% of the input reads were merged if $k_{x,y}$ passed the 0.2 cut-off (for clusters x and y, $k_{x,y}$ is defined as $k_{1,2}=2W/(n_1+n_2)$ where W is the number of read pairs shared between clusters x and y and n_x is the number of reads in cluster x which does not include the other read from its pair within the same cluster); clusters that passed this threshold but which had no similarity hits to each other were not merged. The re-clustering pipeline was run with a minimum of 36 bp overlap for assembly.

Repeat families identified by RepeatExplorer were annotated according to the results of BLAST searches to the Viridiplantae RepeatMasker library, to a database of conserved protein coding domains from transposable elements and to a custom RepeatMasker library comprising all *Fraxinus* sequences (excluding shotgun sequences), all mitochondrial genome sequences from asterids and all plastid genome sequences from Oleaceae available from NCBI (downloaded on 13 February 2014); these BLAST searches were performed as part of the RepeatExplorer pipeline. For repeat families that were not annotated in RepeatExplorer (that is, no significant BLAST hits), or where only very few reads (<2%) had a BLAST hit or separate reads matched different repeat types (that is, inconsistent BLAST hits), contigs were also searched against the nr/nt database in GenBank using BLASTN with an *E* value cut-off[42] of 1×10^{-10}, against the non-redundant database using BLASTX with an *E* value cut-off of 1×10^{-5}, and submitted to Tandem Repeat Finder version 4.07b with default parameters[43]. Annotation of repeat families from the clustering of the 454 and Illumina data was cross-validated by BLAST searching the contigs from each analysis against each other using the BLASTN program in the BLAST+ package (version 2.2.28+) with an *E* value cut-off of 1×10^{-10} and the DUST filter switched off. Any repeat families annotated as plastid or mitochondrial DNA were removed before downstream analyses (see below).

Identification of repeats from the genome assembly. *De novo* identification of repetitive elements from the assembled ash genome sequence was conducted with RepeatModeler version 1.0.7 (http://www.repeatmasker.org/RepeatModeler.html) using RMBlast as the search engine. All unannotated ('unknown') repeat families from the RepeatModeler library were searched against a custom BLAST database of organellar genomes (see above) using BLASTN with an *E* value cutoff of 1×10^{-10} in the BLAST+ package (version 2.2.28+ (ref. 44)). Any repeat families matching plastid or mitochondrial DNA were removed.

To prevent any captured gene fragments within repetitive element families causing the masking of protein coding genes within the ash assembly, the custom repeat libraries were pre-masked using the TAIR10 CDS data set[45] (TAIR10_cds_20101214_updated; downloaded from http://www.arabidopsis.org). First, transposonPSI version 2 (http://transposonpsi.sourceforge.net) was run with the 'nuc' option to identify any transposable-element-related genes within the TAIR10 CDS data set. Sequences with a significant hit to transposable-element-related sequences (*E* value cut-off of 1×10^{-5}) were removed from the TAIR10 CDS file ($n=308$); a further 19 sequences that included the term 'transposon' in their annotation, but which did not have a hit using transposonPSI, were also removed. The filtered TAIR10 CDS data set was used to hard mask the RepeatModeler library, the RepeatExplorer libraries (454 and Illumina) and the library from RepeatMasker using RepeatMasker version 4.0.5 (http://www.repeatmasker.org) with RMblast as the search engine and the following parameter settings: -s –no_is –nolow. The four pre-masked libraries were combined into a single custom repeat library; any repeat families annotated as 'rRNA', 'low-complexity' or 'simple' were removed before combining the libraries. The combined library was then used to identify repetitive elements in the ash genome assembly with RepeatMasker version 4.0.5, using the same parameter settings as above. RepeatMasker results were summarized using ProcessRepeats with the species set to 'eudicotyledons' and using the 'nolow' option.

In addition to the analysis with the combined custom ash repeat library, repeats within the assembly were also annotated by running RepeatMasker separately with each of the four individual repeat libraries with parameter settings as described above. The results were saved in gff format and combined into a single gff file that was then used to inform the process of annotating protein coding genes (see below, 'Gene annotation').

Although the ash genome assembly covers about 99% of the expected genome size based on flow cytometry, about 17% is composed of Ns. Therefore, the repeat content of the genome assembly may be an underestimate of the actual amount of repetitive DNA within the genome. To test whether the about 18% of missing sequence includes additional repetitive elements we analysed the repeat content of individual Illumina reads that do not map to the genome assembly. Quality-trimmed and length-filtered reads from the Illumina short insert libraries (Supplementary Table 1) were mapped to the assembly using the 'Map Reads to Reference' tool in the CLC Genomics Workbench, with both similarity match and length match parameters set to 0.90. Unmapped reads from the 200 bp, 300 bp and 500 bp insert libraries (equating to about 4.8% of all reads from these libraries; see Supplementary Table 1) were searched against the custom library of ash repeats using BLASTN (see Supplementary Table 5) with an *E* value cut-off of 1×10^{-10} and the DUST filter switched off in the BLAST+ package (version 2.2.29+ (ref. 44)).

To test for evidence of the expression of transposable elements, trimmed RNA sequencing reads from five different tissue types (see Supplementary Table 7) were searched against the custom library of ash repeats using BLASTN as described above for the unmapped DNA sequencing reads.

Gene annotation. Protein coding genes were predicted using an evidence-based annotation workflow incorporating protein, cDNA and RNA-seq alignments. Protein sequences from nine species (*Amborella trichopoda*, *A. thaliana*, *Fraxinus pennsylvanica*, *M. guttatus*, *Populus trichocarpa*, *S. lycopersicum*, *Solanum tuberosum*, *V. vinifera* and *Pinus taeda*; Supplementary Table 8) were soft masked for low complexity (segmasker-blast-2.2.30) and aligned to the softmasked (for repeats) final 2451S assembly with exonerate[46] protein2genome version 2.2.0; alignments were filtered at a minimum 60% identity and 60% coverage, except for *F. pennsylvanica*, which were filtered at a minimum of 80% identity and 60% coverage. Publicly available *F. excelsior* expressed sequence tags (12,083 from GenBank) were aligned with GMAP (r20141229)[47] and filtered at a minimum 95% identity and 80% coverage.

RNA-seq reads from the five sequenced RNA samples were filtered for adaptors and quality trimmed, rRNA reads were identified and removed[48] (trim_galore-0.3.3 http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/: -q 20–stringency 5–length 60; sortmerna-1.9: -r 0.25–paired-out). RNA-seq reads were aligned using TopHat (version 2.0.13/Bowtie 2.2.3)[49] and transcript assemblies were generated using three alternative methods: Cufflinks (version 2.2.1)[50], StringTie (version 1.04)[51] and Trinity (genome-guided assembly)[52]. Assembled Trinity transcripts were mapped to the *F. excelsior* assembly using GMAP (r20141229) at 80% coverage and 95% identity. A comprehensive transcriptome assembly was created using Mikado (version 0.8.5, https://github.com/lucventurini/mikado, L. Venturini, manuscript in preparation) on the basis of the GMAP Trinity alignments, Cufflinks and StringTie transcript assemblies. Mikado leverages transcript assemblies generated by multiple methods to improve transcript reconstruction. Loci are first defined across all input assemblies with each assembled transcript scored on the basis of metrics relating to open reading frame and cDNA size, relative position of the open reading frame within the transcript, untranslated region length and presence of multiple open reading frames. The best scoring transcript assembly is then returned along with additional transcripts (splice variants) compatible with the representative transcript.

Protein coding genes were predicted using AUGUSTUS[53] by means of a generalized hidden markov model that took both intrinsic and extrinsic information into account. An AUGUSTUS *ab initio* model was generated on the basis of a subset of cufflinks assembled transcripts identified by similarity support as containing full-length open reading frames. Gene models were predicted using the trained *ab initio* model with the nine sets of cross species protein alignments, RNA-seq junctions (defining introns), and Mikado transcripts as evidence hints. RNA-seq read density was provided as exon hints and repeat information (interspersed repeats) as nonexonpart hints. We generated two alternative AUGUSTUS models by either including or excluding the RNA-seq read depth information. A set of integrated gene models was derived from the two AUGUSTUS runs along with the transcriptome and protein alignments via EVidenceModeller:r20120625 (EVM)[54]. Weights of evidence were manually set following an initial testing and review process as AUGUSTUS predictions with RNA-seq read depth hint, weight 2; AUGUSTUS predictions without RNA-seq read depth hint, weight 1; protein alignment high confidence (greater than 90% coverage, 60% identity) weight 5; protein alignment low confidence (lower than 90% coverage, 60% identity) weight 1; cufflinks transcripts, weight 1; Mikado transcripts, weight 10; RNA-seq splice junctions, weight 1. We identified examples of EVM errors resulting from incomplete genes in the AUGUSTUS gene predictions or non-canonical splicing; to rectify these problems we substituted the EVM model for the overlapping AUGUSTUS model (with RNA-seq read depth hints). To add untranslated region features and alternative splice variants we ran PASA[55] with Mikado transcript assemblies and available *F. excelsior* expressed sequence tags using the corrected EVM models as the reference annotation.

The PASA updated EVM models were further refined by removing gene models that showed no expression support (using all available RNA-seq libraries) or had no support from cross species protein alignments or no BLAST similarity

support with a Viridiplantae (without *F. excelsior*) protein database (<50% BLAST high-scoring segment pair coverage) or where the CDS length was less than 100 bp (retaining those transcripts with ≥50% BLAST high-scoring segment pair coverage). Gene models were also excluded if they aligned with at least 30% similarity and 40% coverage to the TransposonPSI (version 08222010) library (http://transposonpsi.sourceforge.net/) and had at least 40% coverage by the RepeatModeler/RepeatMasker derived interspersed repeats. In addition, gene models that had at least 30% similarity and 60% coverage to the TransposonPSI library or had at least 60% coverage by the RepeatModeler/RepeatMasker derived interspersed repeats were also excluded. The functional annotation of protein coding genes was generated using an in-house pipeline, AnnotF-1.01, which executes and integrates the results from InterProSCAN (version 5) and Blast2GO (version 2.5.0). Completeness of transcript models was classified by Full-lengther Next[56] and coherence in gene length examined by comparison with single copy gene BLAST hits in monkey flower (Extended Data Fig. 1).

Transfer RNA genes were predicted by tRNAscanSE-1.3.1 with eukaryote parameters[57] and rRNAs using RNAmmer-1.2 (ref. 58). miRNA was predicted by BLASTN searches with precursor miRNAs from miRBase[59] 21.0 against the reference genome sequence (BLAST 2.2.30, *E* value 1×10^{-6}) and miRCat[60] using the mature miRNAs from miRBase with default plant parameters, except modifying the flanking window to 200 bp. Putative miRNA precursors from these methods were combined and were folded using RNAfold[61] and mature miRNAs from miRBase were aligned to precursor hairpins using PatMaN[62]. These predictions were checked manually for RNA secondary structure.

Organellar genes were annotated manually using the BLAST tool within the CLC Genomics Workbench version 7.5. Mitochondrial genes were identified using CDS from *M. guttatus*, *Nicotiana tabacum* and *A. thaliana* (all downloaded from NCBI). Plastid genes were identified using CDS from *O. europaea* and *N. tabacum* (both downloaded from NCBI). An *E* value cut-off of 1×10^{-4} was used. Gene and CDS annotations were added manually to the *F. excelsior* organellar scaffolds using the sequence editing tools available within the CLC Genomics Workbench. In the plastid genome, we annotated 72 protein-coding, 7 putative coding (ycf), rRNA and tRNA genes. On the mitochondrial scaffolds, we annotated 37 protein-coding, rRNA and tRNA genes.

Analysis of whole-genome duplications. To examine evidence for past whole-genome duplication, CDS and protein sequences (one transcript per gene) were taken from our ash genome annotation, and downloaded from Phytozome version 10.3 for tomato (*S. lycopersicum*), monkey flower (*M. guttatus*) and grape (*V. vinifera*), the CoGe database for bladderwort (*U. gibba*) and http://coffee-genome.org for coffee (*C. canephora*). For olive (*O. europaea*) we predicted open reading frames from transcriptome data[63] using Transdecoder[52] with all parameters set to defaults (version 2.01, http://transdecoder.github.io). Olive[63] is in the same family as ash (Oleaceae); monkey flower[8] and bladderwort[64] are in the same order as ash (Lamiales); tomato[4] and coffee[65] are in different orders (Solanales and Gentianales, respectively), but like ash in the asterids; and grape[66] is a rosid. An all-against-all comparison using protein sequences was performed on each species separately using BLASTP version 2.2.29, with an *E* value cut-off of 1×10^{-5}. BLAST alignments were further filtered to retain pairs for which the shorter sequence was at least 50% of the longer sequence, and the alignment was at least 50% of the shorter sequence. If one sequence had multiple matches meeting the length and *E* value thresholds, these were grouped into a paralogue group, including any other genes that were associated with the matches (for example, if gene *A* matches gene *B* and gene *C*, and gene *C* also matches gene *D*, then one group of *A*, *B*, *C* and *D* would be formed).

Next, all possible pairs of protein sequences within each group were aligned using muscle version 3.8.31 with default parameters[67]. A nucleotide alignment was generated from the protein alignment using a Python script. Synonymous substitutions were estimated using the codeml program from PAML version 4.8 (ref. 68). The K_s scores within each group were then corrected to remove redundant values; only those representing duplication events within the group were retained (in a group of *n* genes, there are $n-1$ possible duplication events) using the method described in refs 9 and 69. These steps are implemented in a Python script available online: http://github.com/EndymionCooper/KSPlotting.

To examine patterns of conserved synteny, we constructed syntenic dotplots using the SynMap[70] with default parameters (Extended Data Fig. 2). The default uses LAST[71] to perform similarity searches, and DAGchainer[72] to find syntenic regions. By default DAGchainer requires a minimum of 5 aligned gene pairs with no more than 20 genes between neighbouring pairs.

Pairs of genes were categorized as 'local' duplications if they were located on the same chromosome or scaffold and resided within ten genes of each other, and as 'tandem' duplications if they reside directly next to each other. Gene Ontology term enrichment was performed on ash proteins using the BLAST2GO plugin suite of tools within the CLC Genomics Workbench version 8.5. Three separate BLAST searches were run against the RefSeq protein database: first using CDS from all genes as queries, second using CDS from genes involved in whole-genome duplication (excluding locally duplicated genes), and third using CDS from locally duplicated genes (genes located within ten genes of each other). The *E* value cut-off for all BLAST runs was 1×10^{-5}. BLAST results were annotated with Gene Ontology terms using the 'Mapping' and 'Annotation' tools within the BLAST2GO plugin, using default parameters except for Annotation Cutoff = 55 and high-scoring segment pair-hit coverage cutoff = 40. Significantly enriched Gene Ontology terms were identified using the Fisher's exact test tool within the plugin, where the reference set was the Gene Ontology terms for all genes, and a false discovery rate of 0.05 was used.

Analysis of gene families. The OrthoMCL pipeline (version 2.0.9)[73] was used to identify clusters of orthologous and paralogous genes from *F. excelsior* and the following: *Amborella*[74], *Arabidopsis*[75], barrel medic[76], bladderwort[64], coffee[65], grape[66], loblolly pine[77], monkey flower[8], poplar[78] and tomato[4] (Supplementary Table 10). Input proteomes contained a single transcript per gene and were filtered with orthomclFilterFasta to remove any sequences of fewer than ten amino acids in length and/or >20% stop codons. Similar sequences were identified via an all versus all BLASTP search for the 362,741 proteins remaining after filtering. The BLAST search was performed in the BLAST+ package[44] (version 2.2.29+), using an *E* value cut-off of 1×10^{-5}. BLAST results were filtered with orthomclPairs to retain protein pairs that match across at least 50% of the length of the shorter sequence in the pair. Clustering of sequences was performed with mcl[79] (version 14.137) using a setting of 1.5 for the inflation parameter. The output from OrthoMCL was summarized using a custom Perl script to obtain counts of the number of sequences from each species belonging to each group. Venn diagrams for selected taxa were generated using InteractiVenn[80].

European Diversity Panel sequencing. DNA from the 37 European Diversity Panel trees was sequenced at the Earlham Institute on Illumina HiSeq, using paired-end insert sizes between 100 and 700 bp, and a read length of 150 bp. This generated an average of 63.6 million 150 bp reads (10.9× genome coverage) per tree. Filtering and trimming steps reduced this average to 55.3 million reads. An average of 85.8% of these reads per tree mapped to our reference genome. In addition, DNA reads from Danish Tree35 library '3077' were downloaded from the Open Ash Dieback website (http://oadb.tsl.ac.uk); these were 250 bp paired-end reads with an insert size between 200 and 400 bp. Tree35 is given the sample number '38' in all further population analysis.

European Diversity Panel genome-wide SNP calling. The raw reads from the 37 trees in the European Diversity Panel (Supplementary Table 11) were aligned to the reference genome using Bowtie 2.2.5 (ref. 81). The alignments were converted to BAM format and duplicated reads were removed with samtools 1.2 (ref. 82). To assign each read to its corresponding tree, the flag 'rg' was added to each BAM file with picard tools 1.119 (http://broadinstitute.github.io/picard/). SNPs were called with freebayes 1.0.2 (ref. 10) to produce a VCF file. The SNPs with quality less than 300 were filtered with bio-samtools 2.1 (ref. 83). SnpEff 4.1g (ref. 84) was used to predict the effect of the putative SNPs (see Supplementary Table 12). Genic regions were within 5 kbp from a gene model. Amino-acid changes were labelled as missense_variant.

SNP call validation using the KASP platform. To test the reliability of SNP calls in the genome-wide SNP calling, we designed KASP assays for 53 SNPs, which ranged in their level of confidence (see Supplementary Table 13). None of the SNP calls tested by KASP were present in the reduced SNP set used for population genetic analyses. Primers were designed with a modified version of PolyMarker[85] including FAM or HEX tails (FAM tail: 5′-GAAGGTGACCAAGTTCATGCT-3′; HEX tail: 5′-GAAGGTCGGAGTCAACGGATT-3′). The primer mix was prepared as recommended by the manufacturer (46 μl distilled H_2O, 30 μl common primer (100 μM) and 12 μl of each tailed primer (100 μM)) (http://www.lgcgroup.com/services/genotyping). The assays were run on 37 individuals from the European Diversity Panel, in 384-well plates as 4 μl reactions (2-μl template (10–20 ng of DNA), 1.944 μl of V4 2× Kaspar mix and 0.056 μl primer mix). PCR was done with the following protocol: hotstart at 95 °C for 15 min, followed by ten touchdown cycles (95 °C for 20 s; touchdown 65 °C, −1 °C per cycle, 25 s) then followed by 30 cycles of amplification (95 °C for 10 s; 57 °C for 60 s). Fluorescence was detected on a Tecan Safire at ambient temperature. Genotypes were called using Klustercaller software (version 2.22.0.5; LGC Hoddesdon, UK). Four of the individuals did not amplify and were discarded from the analysis. The results of the calls are in Supplementary Data 7.

European Diversity Panel population genetics and history using a reduced set of SNPs. For population structure analyses and effective population size estimation, variants were only called at SNP sites in the genome where all 38 samples had between 5× and 30× coverage. We refer to this as the 'reduced SNP set'.

First, all reads were trimmed in the CLC Genomics Workbench to a minimum quality score of 0.01 (equivalent to Phred quality score of 20), a minimum length

of 50 bp, and were also trimmed of any adaptor and repetitive telomere sequences. Filtered reads were mapped to the reference assembly using the 'Map Reads to Reference' tool in the CLC Genomics Workbench, setting both similarity match and length match parameters to 0.95. Regions with coverage of between 5 and 30 reads in all samples were extracted using the 'Create Mapping Graph', 'Identify Graph Threshold Areas' and 'Calculus Track' tools. These extracted regions totalled 20.6 Mbp (2.3% of the genome).

Variant calling was performed on a read mapping pooled from all samples, using the 'Low Frequency Variant Caller' tool in the CLC Genomics Workbench, with the coverage-restricted regions from the previous step used as a track of target regions. This prevented variants being called where some samples did not have read coverage, and in the organellar scaffolds where the read coverage was very high. The following parameters were changed from default: Ignore positions with coverage above = 1,000, Ignore broken pairs = no, Ignore non-specific matches = Reads, Minimum Coverage = 190 (38 samples with at least 5 reads each should have a combined total coverage of >189), Minimum Count = 10, Minimum Frequency = 5%, Base Quality Filter = Yes, Neighbourhood radius = 5, Minimum Central Quality = 20, Minimum neighbourhood quality = 15, Read Direction Filter = yes, Direction Frequency = 5%. As a result, 529,812 variants were called, comprising 468,237 SNPs, 14,850 equal replacements (where >1 nucleotide is replaced by an equal number of nucleotides), 26,043 deletions, 19,085 insertions and 1,597 unequal replacements (where at least one SNP lies directly beside an indel). The average quality of all reads at these variant positions was 36.2.

To genotype each sample individually at the variant loci called in the previous steps, the 'Identify Known Mutations from sample mappings' tool within the CLC Biomedical Genomics workbench was used. The workflow takes a track of known variants as input (such as those called from the pooled read mapping) and reports the presence, absence, coverage, count and other statistics of each variant locus in the read mapping of another sample (in this case, the read mapping from each of the 38 trees). The 'Identify Candidate Variants' tool was then used to filter variants with a minimum coverage of 5, minimum count of 3 and minimum frequency of 20%. VCF files for each tree were exported from the CLC Workbench and merged into one file using the vcf-merge tool from VCFtools[86]. The merged VCF file was then filtered using vcftools, to remove indels, multi-allelic loci, and loci with a minimum allele frequency < 0.05, with 394,885 SNP loci remaining. This set of high-quality SNPs with comprehensive knowledge of the genotype of every sample was referred to as the 'reduced SNP set' and used for further population analyses.

To visualize similarities and differences among the genomes of the European Diversity Panel, PCA was performed using the SNPRelate version 1.4.2 (ref. 87) package in R version 3.1.2. The filtered VCF file was converted into gds using the snpgdsVCF2GDS command, and was filtered on a linkage disequilibrium value of 0.1 using the snpgdsLDpruning command, leaving 34,607 SNPs. PCA was performed on the pruned set of SNPs using the snpgdsPCA command with default options, and the results of the first three PCs were plotted in R.

To analyse population structure in the European Diversity Panel, scaffolds were selected that contained 10 or more SNPs in the filtered VCF file (8,955 nuclear scaffolds in total). Three different SNPs were selected at random from each of these scaffolds, and placed into three different files in STRUCTURE input format (26,865 SNPs in total, 8,955 in each set). STRUCTURE version 2.3.4 (ref. 88) was run with admixture from $k = 1$ to $k = 20$ for each of the three sets of SNPs, with both BURNIN and NUMREPS set to 100,000. All output results were run through Structure Harvester Web version 0.6.94 (ref. 89), which found $k = 3$ to have the largest Δk value of 32.91 (Extended Data Fig. 3). Next, the three runs of $k = 3$ were used as input into CLUMPP version 1.1.2 (ref. 90) to align the clusters, and samples within each cluster. Aligned results were imported back into STRUCTURE version 2.3.4 to generate *Q* value bar plots. Average *Q* values from the three runs were used to generate a map with pie charts, using Tableau version 9.3 (Tableau, Seattle, USA) with Tableau base-map country outlines. Each section of the pie represented the average *Q* value of the individual belonging to the coloured cluster (Fig. 2b).

To analyse relationships among plastid sequences in plastid haplotype networks, a consensus sequence of the large single copy plastid region was extracted for each of the 38 samples. The sequences were then aligned using the Create Alignment tool in the CLC Genomics Workbench, and the alignment was exported in Phylip format. The alignment was imported into PopArt version 1.7 (http://popart.otago.ac.nz), where a Median Joining network was generated. Results were visualized on a map using Tableau version 9.3 (Fig. 2a) with Tableau base-map country outlines.

We estimated the effective population size history of *F. excelsior* using two complementary methods: the PSMC[14] model estimated the history in the non-recent past, whereas by using linkage disequilibrium, we could estimate the population size more recently. The PSMC model calculated the effective population size using a time to most recent common ancestor approach. The effective population size history was then estimated from the number of recombination events separating segments of constant time to most recent common ancestor. The program PSMC 0.6.5 (ref. 14) took only a diploid consensus sequence as input. To estimate past effective population size, PSMC analysis was used on the reference tree. DNA reads from the 2451S 200, 300 and 500 bp libraries were mapped to the 2451S reference sequence using CLC Genomics Workbench 'Map Reads to Reference' tool (length fraction = 0.95 and similarity fraction = 0.9). The mapping was exported in BAM format, and a consensus sequence was obtained following PSMC recommendations, by using samtools version 0.1.18 'mpileup' command with options -C 50 -A -Q 20 -u, bcftools version 1.1 to convert the BCF file to VCF format, and finally using vcfutils.pl to convert the VCF file to a consensus sequence where the coverage was between 5 and 200. The PSMC program was then run with default parameters except for -p '4+25*2+4+6', with 100 bootstraps. To scale the results, the psmc_plot.pl script was used with default parameters except for the following: -u 7.5e-09 -g 15 -N 0.25 (the mutation rate of *F. excelsior* was unknown, so the substitution rate of 7.5×10^{-9} was taken from a study on *A. thaliana*[91]). Effective population size estimates were then plotted in R version 3.1.2 (Fig. 2f).

Effective population size estimation by linkage disequilibrium in the European Diversity Panel was performed using the program SNeP version 1.1 (ref. 92), which takes genome-wide polymorphism data from several individuals in a population as input. The European Diversity Panel filtered VCF file with the reduced SNP set of 38 trees (the same as used in PCA and STRUCTURE analysis) was converted into Map and Ped files. The third column in the Map file (linkage distance in morgans) was set to zero for all SNPs, as these values were unknown and SNeP calculates this value from each SNP's physical distance. SNeP was then run with a minimum distance between SNPs of 10,000 bp and a maximum of 400,000 bp, with Sved's modifier for recombination rate, and with 50 bins. Estimated effective population sizes were plotted in R (Extended Data Fig. 3c), as well as linkage disequilibrium decay over distance between 100 and 300,000 bp (Fig. 2e).

Simple-sequence repeat analysis. To develop accessible population genetic markers, the repeat masked version 0.4 2451S genome was mined for simple sequence repeat (SSR) sequences (a repeat motif of 2–5 bp in length repeated a minimum of five times) using the QDD version 3.1 pipeline[93]. Downstream QDD version 3.1 pipes screened SSR loci (inclusive of the SSR repeat motif and 200 bp forward and reverse flanking regions) for singleton sequences in an all-against-all BLAST (-task blastn -evalue 1e-40 -lcase_masking -soft_masking true) and designed primer pairs within 200 bp flanking regions using PRIMER3 software[94]. The approximately 31,300 singleton SSR loci identified in the ash genome were screened using RepeatMasker Open-4.0 (http://www.repeatmasker.org) in QDD version 3.1 to eliminate loci that hit known transposable elements in the RepBase Viridiplantae repeat library (http://www.girinst.org), leaving about 28,800 SSR loci. The final primer table output by the QDD version 3.1 pipeline allows selection of the best primer pair design for each SSR loci. To select candidate markers for further development, these primer pairs were filtered according to parameters provided by QDD version 3.1. The selected SSR loci had a: maximum primer alignment score of 5; minimum 20 bp forward and reverse flanking region between SSR and primer sequences; high-quality primer design (defined by QDD pipeline as an absence of homopolymer, nanosatellite and microsatellite sequence in primer and flanking sequences); and minimum number of 7 motif repeats within the SSR sequence. This filtering gave a set of 837 SSR loci, which was screened against the combined custom ash repeat library for v0.5 of the 2451S genome assembly (see above: 'Analysis of repetitive DNA') via a BLASTN search with an *E* value of 1×10^{-10} in the BLAST+ package (version 2.2.31+). Elimination of all sequences with a hit to known repetitive elements left 681 candidate loci. These were compared with the v0.5 assembly via a BLASTN search with an *E* value cut-off of 1×10^{-10}. This returned a set of 664 loci with a unique match to the v0.5 assembly for use as population genetic markers (see Supplementary Data 1).

In silico analysis of allelic diversity (that is, locus polymorphism) of these SSR loci was performed by screening a subset of loci (366) against a variance table composed of insertions and deletions recorded for the European Diversity Panel. Approximately half (48%) of the loci tested were variable among 37 of the re-sequenced genomes (sample 38 not included). Twenty candidate SSR loci with the greatest *in silico* allelic diversity were selected for wet laboratory testing on seven individuals from the European Diversity Panel. Primer pairs with a fluorescent tag on the 5′ end of the forward primer (FAM, HEX or TAM) were used. For singleplex PCR, primer aliquots were used at a concentration of 10 pmol/μl. PCR amplification of target regions was performed in singleplex reactions with a final reaction volume of 10 μl, containing 1 μl genomic DNA, 0.2 μl of each primer (10 pmol/μl), 3.6 μl of RNase free water, and 5 μl of Qiagen Type-it Multiplex PCR Master Mix, in a G-Storm GS2 Multi Block Thermal Cycler. The amplification conditions were as follows: 5 min at 95 °C; 18 cycles of 30 s at 95 °C, 90 s at 62 °C with a 0.5 °C reduction per cycle, 30 s at 72 °C; 20 cycles of 30 s at 95 °C, 1 min 30 s at 51 °C, 30 s at 72 °C; a final extension step of 30 min at 60 °C. PCR samples were diluted to 1:10 with distilled H_2O and run (on an Applied Biosystems 3730xl 96 capillary sequencing instrument with Applied Biosystems GeneScan 400HD Rox

dye size standard). Negative control samples were included for each primer pair PCR reaction mix. Allele calling was performed using GeneMarker version 2.6.4 (http://www.softgenetics.com).

Primer pairs that produced interpretable allele peaks from capillary sequencing of singleplex reactions were arranged into four multiplex primer mixes (containing five primer pairs each) according to PCR product size and fluorescent tag. Multiplex primer mixes were tested on DNA extractions for a further 14 of the 37 trees from the European Diversity Panel. For each multiplex, primer pair mixes were prepared at a final concentration of 10 pmol/μl and amplified via PCR in 10 μl reaction volumes (1 μl genomic DNA, 1 μl primer mix, 3 μl of RNase free water, and 5 μl of Qiagen Type-it Multiplex PCR Master Mix) under the amplification conditions described above. PCR product size range, allele counts, primer design and successful multiplex panels for the 20 wet laboratory tested candidate SSR markers developed for European ash are described in Supplementary Data 1.

Further multiplex primer mixes were tested on 7 trees from the European Diversity Panel for amplification of the longest SSR loci (14 or more repeated motifs). Primer pair mixes were prepared at a final concentration of 10 pmol/μl and amplified via PCR in 8 μl reaction volumes (1 μl genomic DNA from a 1:10 dilution with nuclease free water, 1 μl primer mix, 2 μl of RNase free water, and 4 μl of Qiagen Type-it Multiplex PCR Master Mix.). The amplification conditions were as follows: 5 min at 95 °C; 32 cycles of 30 s at 95 °C, 90 s at 62 °C with a 0.35 °C reduction per cycle, 30 s at 72 °C; a final extension step of 30 min at 60 °C. Amplification was performed in a G-Storm GS2 Multi Block Thermal Cycler. Size fraction analysis of PCR products was performed for two samples of each tested primer multiplex using a 12 sample DNA1000/7500 chip in an Agilent 2100 Bioanalyzer (http://www.genomics.agilent.com). Of the 28 primer pairs tested, 22 successfully amplified across the six primer multiplexes tested (Supplementary Data 1).

Association of transcriptomic markers with reduced susceptibility to ADB in Denmark. Sequence reads for the 'Danish Scored Panel' of 182 Danish ash accessions (as described in ref. 3; sequence reads are available in the European Nucleotide Archive under the study accession number PRJEB10202) were mapped to a reference composed of the complete set of CDS models (including 229 genes identified as possible transposable elements; see above: 'Gene annotation'). This provided transcript abundance estimates for 40,133 CDS models (Supplementary Data 2). Transcript abundance was quantified and normalized as reads per kilobase pairs per million aligned reads (RPKM). After filtering out models exhibiting negligible expression (mean RPKM value of below 0.4), 33,204 CDS models were analysed as potential gene expression markers (GEMs; Supplementary Data 3). SNPs were called by the meta-analysis of alignments (as described in ref. 95) of mRNA-seq reads obtained from each of the 182 accessions. SNP positions were excluded if they did not have a read depth in excess of 20, a base call quality above Q20, missing data below 0.25, and three alleles or fewer. An additional noise threshold was used to reduce the effect of sequencing errors, whereby ambiguous bases were only allowed to be called if both bases were present at 0.15 or above. This resulted in a final set of 394,006 SNPs (Supplementary Data 4) of which 234,519 had minor allele frequencies in excess of 0.05, and all of which were within the CDS models constituting the GEM panel.

The SNP data set for the 182 accessions was entered into the program PSIKO[96] to produce a Q matrix, which was composed of two population clusters. The SNP genotypes, Q matrix and ADB damage scores for these trees[3] were incorporated into a compressed mixed linear model[97] implemented in the GAPIT R package[98], with missing data imputed to the major allele. The kinship matrix used in this analysis was also generated by GAPIT.

GEM associations were calculated by a fixed effect linear model in R with RPKM values and the Q matrix inferred by PSIKO as the explanatory variables and damage score the response variable. Coefficients of correlation (r^2), regression coefficients, constants and significance values were outputted for each regression.

Twenty GEMs were associated with damage scores (Supplementary Data 3). A previous analysis of the gene expression data, based on a simple mRNA transcript reference, identified only 13 GEMs associated with ADB damage in ash[3], with the strongest associations exhibiting higher P values than the present study (best P values 5.31×10^{-12} and 9.83×10^{-13}, respectively). The CDS models for the top three GEMs identified in the present study had very high BLAST similarity to the transcripts for two of the GEMs identified in the previous study. FRAEX38873_v2_000173540.4 ($P = 1.95 \times 10^{-10}$) corresponded with Gene_23247_Predicted_mRNA_scaffold3380 from the previous study, but Gene_19216_Predicted_mRNA_scaffold2427 resolved into two distinct CDS models in the present study (FRAEX38873_v2_000261470.1, $P = 9.83 \times 10^{-13}$ and FRAEX38873_v2_000199610.1, $P = 6.01 \times 10^{-12}$). The qRT–PCR primers designed for the previous analysis[3] were adequate for assaying FRAEX38873_v2_000173540.4 and FRAEX38873_v2_000261470.1 and new primers were designed for FRAEX38873_v2_000199610.1.

Two of the 20 significantly associated GEMs in the present study, FRAEX38873_v2_000048360.1 ($P = 1.77 \times 10^{-9}$) and FRAEX38873_v2_000048340.1 ($P = 3.48 \times 10^{-7}$), did not have high BLAST similarity to GEMS found in the previous study. However, these GEMs were highly similar to a cDNA transcript containing a predictive A/G SNP (termed a cSNP) identified previously, where presence of a G allele was associated with low damage scores. Both of these GEMS contained the 'less susceptible' G variant. A third paralogous gene in this family with the A variant was also found (FRAEX38873_v2_000184430.1), and was not identified as a GEM associated with damage score ($P = 0.02$). The present study therefore resolves this cSNP marker into three paralogous genes, two fixed for a 'less susceptible' G nucleotide, and one a 'susceptible' A nucleotide.

These five GEMs were applied using qRT–PCR, and, in the case of FRAEX38873_v2_000048360.1 and FRAEX38873_v2_000048340.1 RT–PCR, to a small test panel of 58 Danish accessions (henceforth 'Danish Test Panel') to assess their predictive capabilities in a similar way as in ref. 3. Unlike this previous study, however, ratios between the bases of the FRAEX38873_v2_000048360.1 and FRAEX38873_v2_000048340.1 were scored by eye (instead of simply scoring the presence or absence of the 'less susceptible' nucleotide), to estimate levels of gene expression for the 'less susceptible' paralogue, while maintaining the simplicity of the assay. These ratios and the qRT–PCR assays for the other three GEMs were combined into a single predicted damage score for each of the Danish Test Panel, which could then be compared with the observed damage scores for these trees. The combined prediction was correlated with the log mean damage scores for 2013–2014 ($r^2 = 0.25$, $P = 6.9 \times 10^{-5}$) which gave a small improvement in predictive power from the previous analysis ($r^2 = 0.24$, $P < 8.4 \times 10^{-5}$).

Screening of UK *F. excelsior* accessions for markers of reduced susceptibility to ADB. Four markers were selected for predictive marker assays on the basis of this analysis and previous work on the Danish Test Panel of 58 trees[3]. The three GEM markers most highly associated with disease damage were assayed by qRT–PCR using the following primer combinations: FRAEX38873_v2_000261470.1 (GTCGAGGAGGATGGTCAGTCAT, AATCTTGCGGAGGACCTATCG), FRAEX38873_v2_000199610.1 (GGTGAGAGGAAAGGTTCAAATGA, TGCGTTTTGAGAAGGAAACCA), FRAEX38873_v2_000173540.4 (AGGGCAAGGCTTGGAAACAT, TAGGCTTTTTTCTAGCTGCTTGTCA) and *GAPDH* reference (CTGGGATCGCTCTTAGCAAGA, CGATCAAATCAATC ACACGAGAA).

Using RNA extracted from the British Screening Panel, qRT–PCR reactions were performed with SYBR Green fluorescence detection in a qPCR thermal cycler (ViiA™ 7, Applied Biosystems, San Francisco, California) using optical grade 384-well plates, allowing all reactions to be performed simultaneously for each target gene. Each reaction was prepared using 3 μl from a 2 ng/μl dilution of cDNA derived from the RT reaction, 5 μl of SYBR Green PCR Master Mix (Applied Biosystems), 200 nM forward and reverse primers, in a total volume of 10 μl. The cycling conditions were 2 min at 50 °C, 10 min at 95 °C, followed by 40 cycles of 95 °C for 15 s and 60 °C for 1 min with the final dissociation at 95 °C for 15 s, 60 °C for 1 min and 95 °C for 15 s. Three technical replicates were used for quantification analysis. Melting curve analysis was performed to evaluate the presence of non-specific PCR products and primer dimers. The specificity and uniqueness of the primers and the amplicons were verified by amplicon sequencing (GATC Biotech LIGHTrun). The results were exported as raw data, and the LinRegPCR[99] software was used for baseline correction. The resulting means of triplicate N_O values, representing initial concentrations of a target and reference gene, were used to analyse gene expression. For each marker, the set of qRT–PCR quantifications were standardized and rescaled to better emulate the range of RPKM values observed in the original association panel, and then predicted damage scores generated using the regression coefficient and constant from the GEM associations.

An additional GEM marker was assayed as a cSNP by PCR using 1 μl undiluted cDNA, 11.5 μl Thermo Scientific Fermentas PCR Master Mix (2×), 200 nM forward (GGTTTCTCTTCTGCAGCGAG) and reverse (TCCATGA TCATCTTGCTGAG) primers in a total volume of 25 μl. The touchdown PCR was performed in using a BIORAD Tetrad PCR machine with the following cycling conditions: 5 min at 94 °C, followed by 15 cycles of 94 °C for 30 s, 63 °C for 30 s −1 °C per cycle, 72 °C for 1 min, and 30 cycles of 94 °C for 30 s, 53 °C for 30 s, 72 °C for 1 min and a final elongation step at 72 °C for 7 min.

Sanger sequences obtained using the forward primer co-amplify GEM FRAEX38873_v2_000048360.1, which is highly associated with ADB disease damage, and another member of the gene family that is not. Owing to a polymorphism between the two (at position 203 of the CDS model mentioned above), the relative abundance of the G nucleotide found in the highly associated GEM could be scored by eye relative to the A nucleotide found in the other paralogue as a cSNP. Previously[3], this marker was scored in the Danish Test Panel as the presence or absence of a G nucleotide at this position, but predictions using this method did not incorporate the dynamic range of the gene expression observed. So, for this

analysis, G:A peak height ratios were approximated directly from the sequence chromatograms using Softgenetics Mutation Surveyor software for the British Screening Panel and the Danish Test Panel. These ratios were then standardized and rescaled to the RPKM values for FRAEX38873_v2_000048360.1 to predict damage scores as before.

Combined predictions were made by ranking and standardizing the individual predictions for all four markers, and then calculating the mean rank score for each individual tree (Supplementary Data 6). Combined predictions were calculated for the Danish Test Panel and compared with the observed ADB damage scores to ensure that the assay was predictive (Fig. 3).

The four assays were applied in the same way to analyse a panel of 130 accessions originating from across the UK range of *F. excelsior* ('British Screening Panel'). Strikingly, when assayed by RT–PCR, expression of the 'G' variant paralogues was seen at much higher frequency in the British Screening Panel than in the Danish panels and the mean G:A ratio across the British Screening Panel was 0.67 compared with a mean of 0.03 observed in the Danish Test Panel. Likewise, the gene expression estimates for the British Screening Panel exhibited wider ranges and were more favourable in terms of their expected effect on damage scores. The qRT–PCR results for the GEMs negatively correlated with disease damage (FRAEX38873_v2_000261470.1 and FRAEX38873_v2_000199610.1) exhibited higher mean expression in the UK (0.1 ± 0.11 and 0.12 ± 0.14) versus the Danish Test Panel (0.09 ± 0.08, 0.12 ± 0.11), and the positively correlated FRAEX38873_v2_000173540.4 was on average expressed at a lower level in the British Screening Panel (0.48 ± 0.26) than the Danish Test Panel (0.59 ± 0.17). As expected, this translated to lower combined predictions for ADB damage in the British Screening Panel. Only 9% of the Danish Test Panel accessions were predicted to have a low damage score (defined as 25% canopy damage or less) compared with 25% of the British Screening Panel (Fig. 3).

Analysis of predictive genes. To predict the susceptibility of the reference tree 2451S to ADB, we calculated RPKM values for the five GEM marker CDS models (FRAEX38873_v2_000173540.4, FRAEX38873_v2_000048340.1, FRAEX38873_v2_000048360.1, FRAEX38873_v2_000261470.1 and FRAEX38873_v2_000199610.1) from leaf transcriptome read data. We also did this for each of the trees in the Danish Scoring Panel, and the average of these predictions was taken to provide combined predictions. The top and bottom quartiles from the distribution of predicted scores, which represent the trees with the most susceptible and least susceptible gene expression patterns at these five loci, were then correlated with the RPKM values for the genome sequenced tree 2451S (Extended Data Fig. 4).

RPKM data were also generated for four tissue types: leaf, flower, cambium and root, of the parent of sequenced tree 2451S by mapping raw reads to the CDS reference as before. RPKM data for the 20 CDS models found to be significantly associated with susceptibility to ADB in the GEM analysis were selected and compared for the four tissue types.

The five CDS models represented in the ADB susceptibility predictions were translated using the standard codon usage table and were searched against the non-redundant database in GenBank using BLASTP with default settings to identify top hits to protein sequences in *A. thaliana*: FRAEX38873_v2_000199610.1 and FRAEX38873_v2_000261470.1 show high similarity to AGAMOUS-LIKE 42/FOREVER YOUNG FLOWER (AGL42/FYF; AT5G62165); FRAEX38873_v2_000173540.4, FRAEX38873_v2_000048340.1 and FRAEX38873_v2_000048360.1 have top hits to SHORT VEGETATIVE PHASE/AGAMOUS-LIKE 22 (SVP/AGL22; AT2G22540). Both AGL42/FYF and SVP/AGL22 are encoded by type II MADS-box genes[16]. To find potential orthologues from other species, we examined the results of the OrthoMCL analysis for clusters containing AGL42/FYF and SVP/AGL22; all sequences from these clusters were extracted and added to the appropriate *F. excelsior* sequences to create two data sets, one of AGL42/FYF-like sequences and one of SVP/AGL22-like sequences. To ensure adequate representation of putative orthologues, we further expanded these data sets to include sequences from the OrthoMCL clusters containing *A. thaliana* proteins from closely related MADS lineages, as identified by previous phylogenetic analyses of type II MADS-box sequences[16,17].

Preliminary phylogenetic analysis of these data sets revealed that, despite showing high sequence similarity in BLAST searches, FRAEX38873_v2_000048340.1 and FRAEX38873_v2_000048360.1 do not fall within the clade containing SVP/AGL22 and similar *A. thaliana* sequences. Therefore, to identify potentially more closely related sequences we performed a BLASTP search of FRAEX38873_v2_000048340.1 and FRAEX38873_v2_000048360.1 against the complete set of 362,741 protein sequences used for the OrthoMCL analysis (see Supplementary Table 10), using the BLAST+ package[44] (version 2.2.31+) with an *E* value cut-off of 1×10^{-5} (FRAEX38873_v2_000048340.1 and FRAEX38873_v2_000048360.1 were not included in the OrthoMCL analysis because they were flagged as putative transposable-element-related genes during annotation). This identified several highly similar sequences from other species with better ranking BLAST hits than those to the *A. thaliana* proteins. These sequences belong to a single OrthoMCL cluster, and include a tomato (*S. lycopersicum*) sequence from the apparent orthologue of the potato (*S. tuberosum*) *StMADS11* gene; all sequences from this cluster were added to the SVP/AGL22-like data set, along with the potato StMADS11 protein (GenBank accession number ACH53556.1).

Sequences for both data sets were aligned using M-Coffee[100], via the T-Coffee web server (http://www.tcoffee.org; last accessed 7 December 2016) with the following parameter settings: Mpcma_msa Mmafft_msa Mclustalw_msa Mdialigntx_msa Mpoa_msa Mmuscle_msa Mprobcons_msa Mt_coffee_msa -output = score_html clustalw_aln fasta_aln score_ascii phylip -tree -maxnseq = 150 -maxlen = 2500 -case = upper -seqnos = on -outorder = input -run_name = result -multi_core = 4 -quiet = stdout. Positions in the alignments with consensus scores of <6 from M-Coffee were removed; filtered alignments were then run through the TCS tool[101] via the T-Coffee web server and any positions with a reliability score of <6 were removed. Recombination was tested for in the filtered alignments using GARD[102]. Analyses were run via the Datamonkey server (http://www.datamonkey.org; last accessed 1 June 2016) under the best-fit model of evolution (selected with the corrected Akaike's information criterion[103]) with β–Γ rate variation and three rate classes. No breakpoints with significant topological incongruence at $P \leq 0.05$ were detected for either data set. Phylogenetic analysis of each data set was conducted using Bayesian inference in MrBayes and maximum likelihood in RAxML; input alignments are provided in Supplementary Data 8. MrBayes (version 3.2.5 (ref. 104)) was run using the mixed amino acid model, to allow models of protein sequence evolution to be fit automatically across the alignments; the following parameter settings were used for each data set: prset aamodellpr = mixed, mcmc nruns = 2, nchains = 4, ngen = 1000000, samplefreq = 1000. Parameter values from both runs for each data set were viewed in TRACER version 1.6 (http://beast.bio.ed.ac.uk/Tracer) to confirm that effective sample sizes of >200 had been obtained for each parameter and stationarity reached. Trees sampled during the first 100,000 generations of each run were discarded as the burn-in; trees and parameter values were summarized in MrBayes using the sumt and sump commands. RAxML (version 8.2.8 (ref. 105)) was run using the option to automatically determine the best protein substitution model, with 1,000 replicates of the rapid bootstrap algorithm; parameter settings were as follows: raxmlHPC -f a -x 13102 -p 29503 -# 1000 -m PROTGAMMAAUTO.

The phylogenetic analysis suggested that FRAEX38873_v2_000173540.4 is a likely orthologue of the *A. thaliana SVP*/*AGL22* gene, or possibly *AGL24*, whereas FRAEX38873_v2_000048340.1 and FRAEX38873_v2_000048360.1 appear orthologous to the potato *StMADS11* gene (Extended Data Fig. 5). These all belong to the SVP/StMADS11 group[16] of type II MADS-box genes. FRAEX38873_v2_000261470.1 and FRAEX38873_v2_000199610.1 cluster with the *A. thaliana* SUPPRESSOR of OVEREXPRESSION of CONSTANS 1(SOC1)-like proteins AGL42, AGL71 and AGL72 (Extended Data Fig. 5). The two other major clades within the phylogenetic tree include the AGL20/SOC1 protein and the AG14 and AGL19 proteins (Extended Data Fig. 5); together, the AGL42/AGL71/AGL72-, AL20- and AGL14/AGL19-containing clades are known as the SOC1/TM3 group of type II MADS-box proteins[16,17].

In *A. thaliana*, *AGL42*, *AGL71* and *AGL72* have redundant functions in controlling flowering time and appear to be regulated by *AGL20*/*SOC1* (ref. 20). In turn, *AGL20*/*SOC1* is regulated by both *AGL22*/*SVP* and *AGL24* (refs 18, 19), which are floral meristem identity genes with redundant functions during early stages of flower development[21]. The *StMADS11* gene does not appear to have a direct orthologue in *A. thaliana*, but in potato (*S. tuberosum*) *StMADS11* is expressed in vegetative tissues[106]. Despite their well-known roles in floral regulation, *SVP*/*StMADS11* and *SOC1*/*TM3* proteins are likely to have wider functions. In *A. thaliana*, it is suggested that *AGL22*/*SVP* is also required for age-related resistance, which gives older tissues of plants enhanced pathogen tolerance or resistance[24]. The *B. rapa BrMADS44* gene, which appears orthologous to *AGL42*, shows differential expression in response to cold and drought stress; some *B. rapa* genes belonging to the SVP/StMADS11 clade are also differentially expressed in response to these stresses, indicating a potential role in stress resistance[22]. Furthermore, many genes involved in regulation of flowering time in *A. thaliana* are involved in controlling phenology in perennial trees species and genes belonging to the SVP/StMADS11 clade have potential roles in growth cessation, bud set and dormancy[23].

Metabolomic profiling. To understand whether trees with low and high susceptibility vary in their metabolite profiles as well as their transcriptomes, we undertook untargeted metabolite profiling on a subset of the Danish Test Panel. Untargeted metabolomics has not previously been applied to natural populations but has the potential to identify small molecules (or small-molecule associations) that directly contribute to tolerance or resistance. We compared triplicate samples from five low-susceptibility Danish trees (R-14164C, R-14184A, R-14193A, R-14198B, R-14181) and five high-susceptibility trees (R-14169, R-14127, R-14156 R-14120, 25UTaps).

Three leaflets from each triplicate sample were freeze dried and gently crushed to mix tissue. Approximately 100–150 mg was ground to a fine powder using a TissueLyser (Qiagen), and 10 mg was extracted in 400 µl 80% MeOH containing d5-IAA internal standard at 2.5 ng/ml ([$^2H_5$]indole-3-acetic acid; OlChemIm, Czech Republic), centrifuged (10,000 *g*, 4 °C, 10 min) and the pellet re-extracted in 80% MeOH. The pooled supernatants were filtered through a 0.2 µm syringe filter (Phenomenex, UK).

These leaf extracts (5 µl) were analysed using a Polaris C18 1.8 µm, 2.1 mm × 250 mm reverse-phase analytical column (Agilent Technologies, Palo Alto, California, USA) and samples resolved on an Agilent 1200 series Rapid Resolution HPLC system coupled to a quadrupole time-of-flight QToF 6520 mass spectrometer (Agilent Technologies, Palo Alto, California, USA). Buffers were as follows: positive ion mode; mobile phase A (5% acetonitrile, 0.1% formic acid), mobile phase B (95% acetonitrile with 0.1% formic acid); negative ion mode; mobile phase A (5% acetonitrile with 1 mM ammonium fluoride), mobile phase B (95% acetonitrile). The following gradient was used: 0–10 min, 0% B; 10–30 min, 0–100% B; 30–40 min, 100% B. The flow rate was 0.25 ml/min and the column temperature was held at 35 °C throughout. The source conditions for electrospray ionization were as follows: gas temperature was 325 °C with a drying gas flow rate of 9 l/min and a nebulizer pressure of 35 pounds per square inch gauge. The capillary voltage was 3.5 kV in both positive and negative ion mode. The fragmentor voltage was 115 V and skimmer 70 V. Scanning was performed using the autoMS/MS function at four scans per second for precursor ion surveying and three scans per second for MS/MS with a sloped collision energy of 3.5 V per 100 Da with an offset of 5 V.

Positive and negative ion data were converted into mzData using the export option in Agilent MassHunter. Peak identification and alignment was performed using the Bioconductor R package xcms[107] and features were detected using the centWave method[108] for high-resolution liquid chromatography/mass spectrometry data in centroid mode at 30 p.p.m. Changes from the default parameters were mzdiff = 0.01, peakwidth = 10-80, noise = 1000, prefilter = 3,500. Peaks were matched across samples using the density method with a bw = 5 and mzwid = 0.025 and retention time correlated using the obiwarp algorithm with profStep = 0.5. Missing peak data were filled in the peaklists generated from the ADB low-susceptibility ash leaf samples compared with the peaklists generated from the ADB susceptible leaves. The resulting peaklists were annotated using the Bioconductor R package, CAMERA[109]. The peaks were grouped using 0.05% of the width of the full width at half maximum, and groups correlated using a *P* value of 0.05 and calculating correlation inside and across samples. Isotopes and adducts were annotated using a 10 p.p.m. error.

Statistical analysis and modelling was performed using MetaboAnalyst version 3.0 with the following parameters. Missing values were replaced using a KNN missing value estimation. Data were filtered (40%) to remove non-informative variables using the interquartile range. Samples were normalized using the internal standard d5-IAA (POS: M181T1448; NEG: M179T1382). Data were auto-scaled.

Peaks from the three replicates were aligned with xcms for both positive and negative mode and features tested for practical significance to determine the differences between the tolerant and susceptible genotypes. In addition, PLS-DA was performed using MetaboAnalyst, allowing the discrimination of tolerant and susceptible genotypes on the basis of their metabolic profiles (Fig. 4a).

The individual features (putative metabolites) that contributed to the separation between the different classes were further characterized. We first applied a range of univariate and multivariate statistical tests to determine the importance of these features. This included variable influence on the projection (VIP) values derived from PLS-DA scores, practical significance, *t*-test, *P* value, Benjamini and Hochberg false discovery rate *P* value, effect size and Random Forest analysis, and MS/MS fragmentation network analysis. For example, using Random Forest, significant features were ranked by mean decrease in classification accuracy with 14 out of 15 susceptible samples (out-of-bag error: 0.033; class error 0.07) and 15 out of 15 tolerant samples correctly classified.

For all further analyses we chose to use statistical and practical significance (Response Screening, JMP version 12) to identify features with a practical significance for identification. A combination of *k*-means clustering was used to group features by patterns of abundance and by retention time. This enabled the clustering of base peaks with their associated isotopes and adducts. Product ions were identified using MS/MS data in Agilent MassHunter Qualitative Analysis version 4.

Identification was not possible for those features with no fragmentation, or lacking significant supporting adducts. Many features of interest were identified but require further work to provide confident attributions, while some features did not provide fragmentation patterns. We thus restricted further identification and characterization to a highly discriminatory class of compounds of the iridoid glycosides and predominantly compounds previously recorded in Oleaceae, summarized in Extended Data Figs 6–9 and Supplementary Data 9. We validated these identifications using three methods: MS/MS fragmentation networking (Fig. 4c), MS/MS mirror plot (Extended Data Fig. 6) and accurate mass MS/MS product ion structure correlation (Extended Data Fig. 7). The MS/MS fragmentation network was generated after extracting the *m*/*z* values of the MS/MS product ions from the discriminatory features using MassHunter Qualitative Analysis Version 4 and visualized using Cytoscape, indicating product ion masses that had been previously reported from fragmentation of iridoid glycosides[110]. Further validation was performed through a mirror plot comparing the MS/MS spectra of four features (N_2–N_5) detected in negative mode with an electrospray ionization-time of flight/ion trap-mass spectrometry (ESI-TOF/IT-MS) spectrum of elenolic acid glucoside taken from the literature[111]. Finally, the accurate masses of MS/MS product ions from four discriminatory features identified in negative mode (N_1–N_4) were correlated with the structure of the putatively identified compound using MassHunter Molecular Structure Correlator (Agilent).

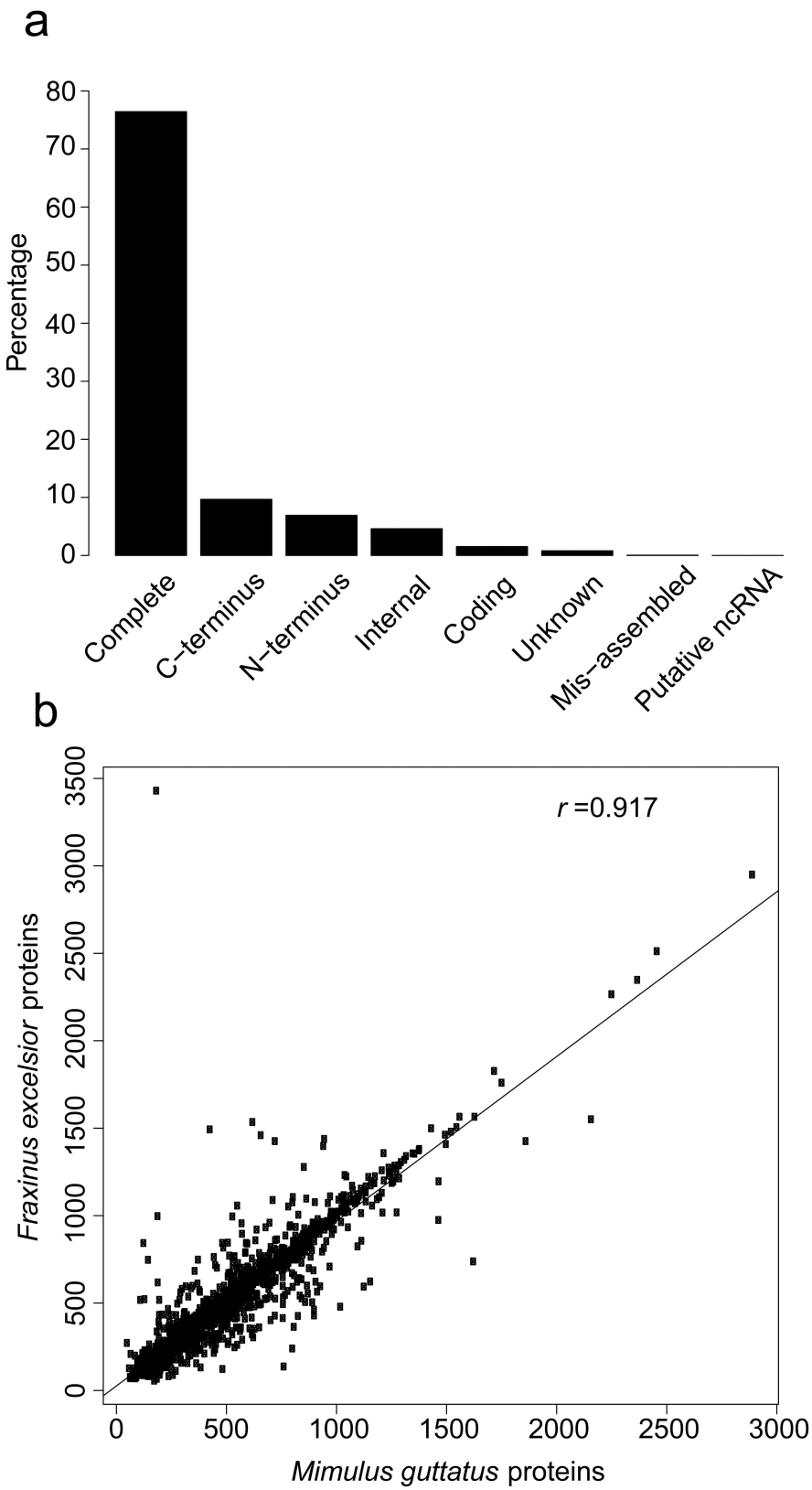
A timeline for the project may be found in Supplementary Table 14.

URL. Genome website: http://www.ashgenome.org.

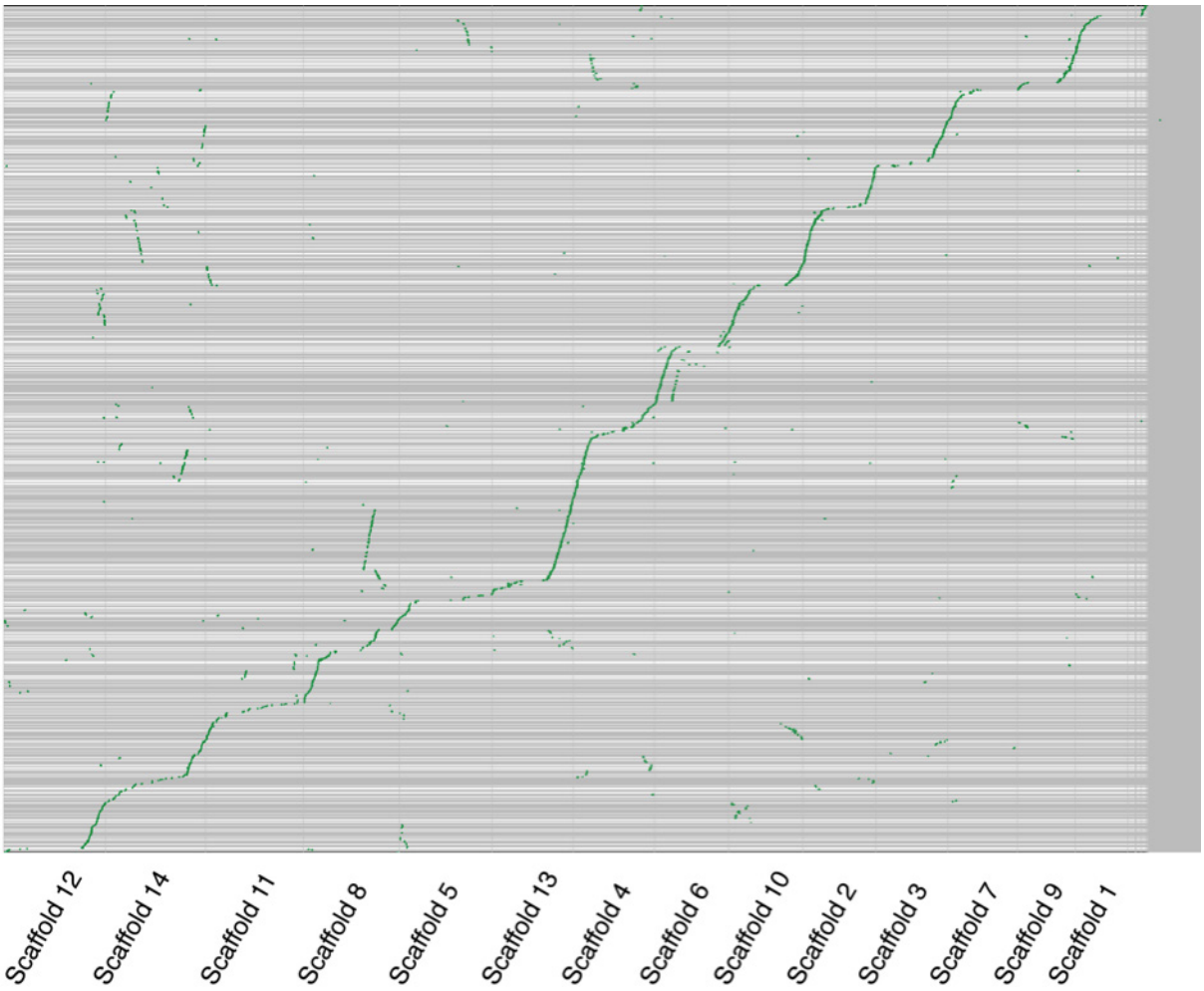
Data availability. The reference tree is growing at Earth Trust with accession number 2451S. Trimmed DNA and RNA reads and the final assembly for the 2451S genome sequence, as well as RNA reads for parent tree and raw reads and consensus read mappings of the European diversity panel trees, have been deposited in European Nucleotide Archive under project accession code PRJEB4958 (http://www.ebi.ac.uk/ena/data/view/PRJEB4958). Metabolomic data that support the findings of this study have been deposited in MetaboLights under accession code MTBLS372 (http://www.ebi.ac.uk/metabolights/MTBLS372). All other data are available from the corresponding author upon reasonable request.

31. FRAXIGEN. *Ash Species in Europe: Biological Characteristics and Practical Guidelines for Sustainable Use* (Oxford Forestry Institute, Univ. Oxford, 2005).
32. Doyle, J. J. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochem. Bull.* **19,** 11–15 (1987).
33. Obermayer, R., Leitch, I. J., Hanson, L. & Bennett, M. D. Nuclear DNA C-values in 30 species double the familial representation in pteridophytes. *Ann. Bot.* **90,** 209–217 (2002).
34. Doležel, J., Greilhuber, J. & Suda, J. Estimation of nuclear DNA content in plants using flow cytometry. *Nature Protocols* **2,** 2233–2244 (2007).
35. Magoč, T. & Salzberg, S. L. FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* **27,** 2957–2963 (2011).
36. Leggett, R. M., Clavijo, B. J., Clissold, L., Clark, M. D. & Caccamo, M. NextClip: an analysis and read preparation tool for Nextera Long Mate Pair libraries. *Bioinformatics* **30,** 566–568 (2014).
37. Boetzer, M., Henkel, C. V., Jansen, H. J., Butler, D. & Pirovano, W. Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics* **27,** 578–579 (2011).
38. English, A. C. *et al.* Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology. *PLoS ONE* **7,** e47768 (2012).
39. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of *k*-mers. *Bioinformatics* **27,** 764–770 (2011).
40. Gomez-Alvarez, V., Teal, T. K. & Schmidt, T. M. Systematic artifacts in metagenomes from complex microbial communities. *ISME J.* **3,** 1314–1317 (2009).
41. Novák, P., Neumann, P., Pech, J., Steinhaisl, J. & Macas, J. RepeatExplorer: a Galaxy-based web server for genome-wide characterization of eukaryotic repetitive elements from next-generation sequence reads. *Bioinformatics* **29,** 792–793 (2013).
42. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215,** 403–410 (1990).
43. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27,** 573–580 (1999).
44. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10,** 421 (2009).
45. Lamesch, P. *et al.* The *Arabidopsis* Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Res.* **40,** D1202–D1210 (2012).
46. Slater, G. S. C. & Birney, E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6,** 31 (2005).
47. Wu, T. D. & Watanabe, C. K. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* **21,** 1859–1875 (2005).
48. Kopylova, E., Noé, L. & Touzet, H. SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. *Bioinformatics* **28,** 3211–3217 (2012).
49. Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-seq. *Bioinformatics* **25,** 1105–1111 (2009).
50. Trapnell, C. *et al.* Transcript assembly and quantification by RNA-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnol.* **28,** 511–515 (2010).
51. Pertea, M. *et al.* StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature Biotechnol.* **33,** 290–295 (2015).
52. Haas, B. J. *et al. De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature Protocols* **8,** 1494–1512 (2013).

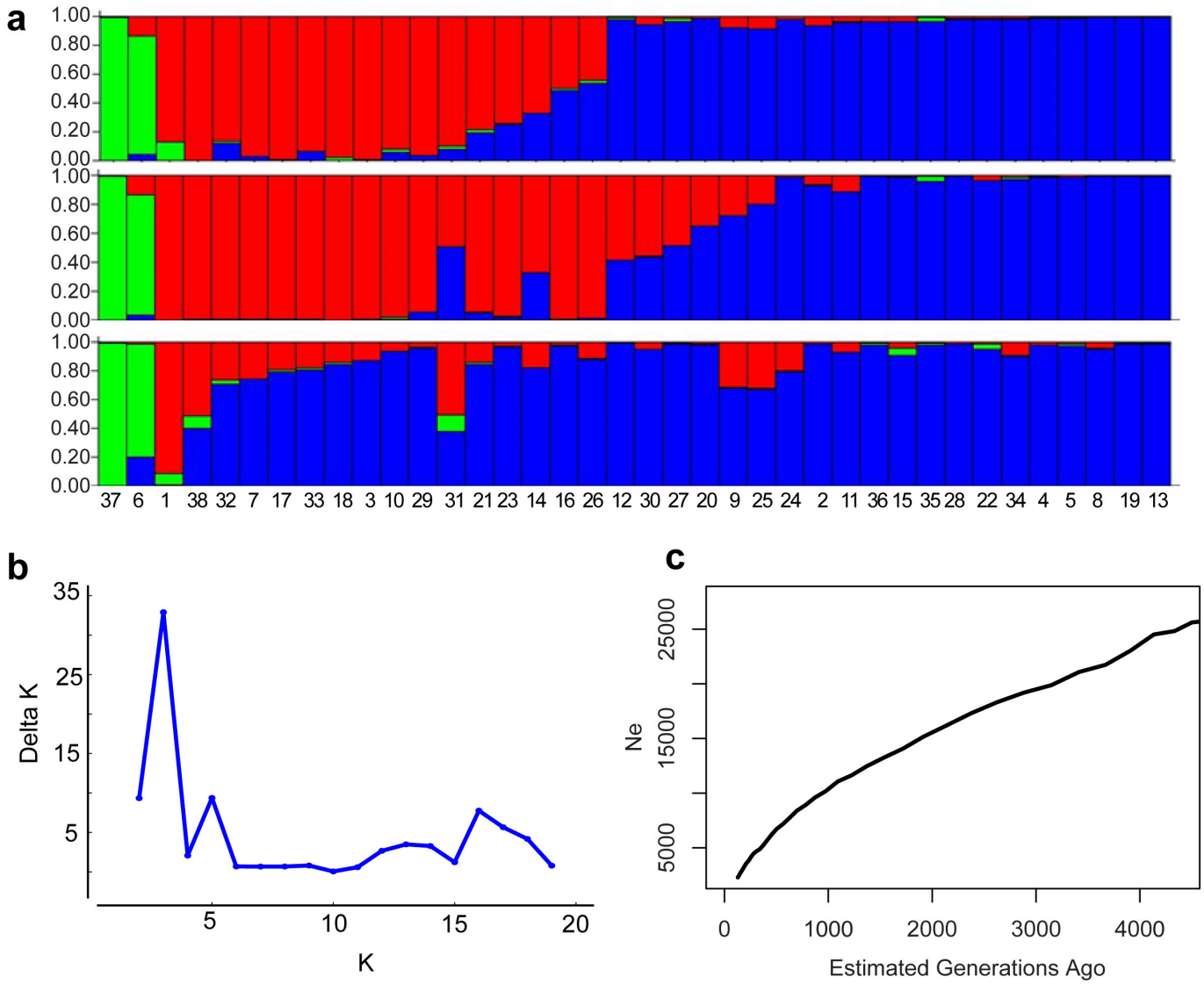
53. Stanke, M., Steinkamp, R., Waack, S. & Morgenstern, B. AUGUSTUS: a web server for gene finding in eukaryotes. *Nucleic Acids Res.* **32,** W309–W312 (2004).
54. Haas, B. J. *et al.* Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol.* **9,** R7 (2008).
55. Haas, B. J. *et al.* Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.* **31,** 5654–5666 (2003).
56. Lara, A. J. *et al.* in *Innovations in Hybrid Intelligent Systems* 361–368 (Springer, 2007).
57. Lowe, T. M. & Eddy, S. R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* **25,** 955–964 (1997).
58. Lagesen, K. *et al.* RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res.* **35,** 3100–3108 (2007).
59. Kozomara, A. & Griffiths-Jones, S. miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res.* **42,** D68–D73 (2014).
60. Stocks, M. B. *et al.* The UEA sRNA workbench: a suite of tools for analysing and visualizing next generation sequencing microRNA and small RNA datasets. *Bioinformatics* **28,** 2059–2061 (2012).
61. Lorenz, R. *et al.* ViennaRNA Package 2.0. *Algorithms Mol. Biol.* **6,** 26 (2011).
62. Prüfer, K. *et al.* PatMaN: rapid alignment of short sequences to large databases. *Bioinformatics* **24,** 1530–1531 (2008).
63. Muñoz-Mérida, A. *et al. De novo* assembly and functional annotation of the olive (*Olea europaea*) transcriptome. *DNA Res.* **20,** 93–108 (2013).
64. Ibarra-Laclette, E. *et al.* Architecture and evolution of a minute plant genome. *Nature* **498,** 94–98 (2013).
65. Denoeud, F. *et al.* The coffee genome provides insight into the convergent evolution of caffeine biosynthesis. *Science* **345,** 1181–1184 (2014).
66. Jaillon, O. *et al.* The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* **449,** 463–467 (2007).
67. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32,** 1792–1797 (2004).
68. Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24,** 1586–1591 (2007).
69. Maere, S. *et al.* Modeling gene and genome duplications in eukaryotes. *Proc. Natl Acad. Sci. USA* **102,** 5454–5459 (2005).
70. Lyons, E., Pedersen, B., Kane, J. & Freeling, M. The value of nonmodel genomes and an example using SynMap Within CoGe to dissect the hexaploidy that predates the rosids. Trop. *Plant Biol.* **1,** 181–190 (2008).
71. Kiełbasa, S. M., Wan, R., Sato, K., Horton, P. & Frith, M. C. Adaptive seeds tame genomic sequence comparison. *Genome Res.* **21,** 487–493 (2011).
72. Haas, B. J., Delcher, A. L., Wortman, J. R. & Salzberg, S. L. DAGchainer: a tool for mining segmental genome duplications and synteny. *Bioinformatics* **20,** 3643–3646 (2004).
73. Li, L., Stoeckert, C. J., Jr & Roos, D. S. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.* **13,** 2178–2189 (2003).
74. Amborella Genome Project. The *Amborella* genome and the evolution of flowering plants. *Science* **342,** 1241089 (2013).
75. The Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408,** 796–815 (2000).
76. Young, N. D. *et al.* The *Medicago* genome provides insight into the evolution of rhizobial symbioses. *Nature* **480,** 520–524 (2011).
77. Zimin, A. *et al.* Sequencing and assembly of the 22-gb loblolly pine genome. *Genetics* **196,** 875–890 (2014).
78. Tuskan, G. A. *et al.* The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science* **313,** 1596–1604 (2006).
79. Enright, A. J., Van Dongen, S. & Ouzounis, C. A. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* **30,** 1575–1584 (2002).
80. Heberle, H., Meirelles, G. V., da Silva, F. R., Telles, G. P. & Minghim, R. InteractiVenn: a web-based tool for the analysis of sets through Venn diagrams. *BMC Bioinformatics* **16,** 169 (2015).
81. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9,** 357–359 (2012).
82. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).
83. Etherington, G. J., Ramirez-Gonzalez, R. H. & MacLean, D. bio-samtools 2: a package for analysis and visualization of sequence and alignment data with SAMtools in Ruby. *Bioinformatics* **31,** 2565–2567 (2015).
84. Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly (Austin)* **6,** 80–92 (2012).
85. Ramirez-Gonzalez, R. H., Uauy, C. & Caccamo, M. PolyMarker: A fast polyploid primer design pipeline. *Bioinformatics* **31,** 2038–2039 (2015).
86. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27,** 2156–2158 (2011).
87. Zheng, X. *et al.* A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics* **28,** 3326–3328 (2012).
88. Hubisz, M. J., Falush, D., Stephens, M. & Pritchard, J. K. Inferring weak population structure with the assistance of sample group information. *Mol. Ecol. Resour.* **9,** 1322–1332 (2009).
89. Earl, D. A. & vonHoldt, B. M. STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conserv. Genet. Resour.* **4,** 359–361 (2012).
90. Jakobsson, M. & Rosenberg, N. A. CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics* **23,** 1801–1806 (2007).
91. Buschiazzo, E., Ritland, C., Bohlmann, J. & Ritland, K. Slow but not low: genomic comparisons reveal slower evolutionary rate and higher dN/dS in conifers compared to angiosperms. *BMC Evol. Biol.* **12,** 8 (2012).
92. Barbato, M., Orozco-terWengel, P., Tapio, M. & Bruford, M. W. SNeP: a tool to estimate trends in recent effective population size trajectories using genome-wide SNP data. *Front. Genet.* **6,** 109 (2015).
93. Meglécz, E. *et al.* QDD version 3.1: a user-friendly computer program for microsatellite selection and primer design revisited: experimental validation of variables determining genotyping success rate. *Mol. Ecol. Resour.* **14,** 1302–1313 (2014).
94. Rozen, S. & Skaletsky, H. Primer3 on the WWW for general users and for biologist programmers. *Methods Mol. Biol.* **132,** 365–386 (2000).
95. Bancroft, I. *et al.* Dissecting the genome of the polyploid crop oilseed rape by transcriptome sequencing. *Nature Biotechnol.* **29,** 762–766 (2011).
96. Popescu, A.-A., Harper, A. L., Trick, M., Bancroft, I. & Huber, K. T. A novel and fast approach for population structure inference using kernel-PCA and optimization. *Genetics* **198,** 1421–1431 (2014).
97. Zhang, Z. *et al.* Mixed linear model approach adapted for genome-wide association studies. *Nature Genet.* **42,** 355–360 (2010).
98. Lipka, A. E. *et al.* GAPIT: genome association and prediction integrated tool. *Bioinformatics* **28,** 2397–2399 (2012).
99. Ruijter, J. M. *et al.* Amplification efficiency: linking baseline and bias in the analysis of quantitative PCR data. *Nucleic Acids Res.* **37,** e45 (2009).
100. Di Tommaso, P. *et al.* T-Coffee: a web server for the multiple sequence alignment of protein and RNA sequences using structural information and homology extension. *Nucleic Acids Res.* **39,** W13–W17 (2011).
101. Chang, J.-M., Di Tommaso, P. & Notredame, C. TCS: a new multiple sequence alignment reliability measure to estimate alignment accuracy and improve phylogenetic tree reconstruction. *Mol. Biol. Evol.* **31,** 1625–1637 (2014).
102. Kosakovsky Pond, S. L., Posada, D., Gravenor, M. B., Woelk, C. H. & Frost, S. D. W. GARD: a genetic algorithm for recombination detection. *Bioinformatics* **22,** 3096–3098 (2006).
103. Sugiura, N. Further analysis of the data by Akaike's information criterion and the finite corrections. *Commun. Stat. Theory Methods* **7,** 13–26 (1978).
104. Ronquist, F. *et al.* MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst. Biol.* **61,** 539–542 (2012).
105. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30,** 1312–1313 (2014).
106. Carmona, M. J., Ortega, N. & Garcia-Maroto, F. Isolation and molecular characterization of a new vegetative MADS-box gene from *Solanum tuberosum* L. *Planta* **207,** 181–188 (1998).
107. Smith, C. A., Want, E. J., O'Maille, G., Abagyan, R. & Siuzdak, G. XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Anal. Chem.* **78,** 779–787 (2006).
108. Tautenhahn, R., Böttcher, C. & Neumann, S. Highly sensitive feature detection for high resolution LC/MS. *BMC Bioinformatics* **9,** 504 (2008).
109. Kuhl, C., Tautenhahn, R., Böttcher, C., Larson, T. R. & Neumann, S. CAMERA: an integrated strategy for compound spectra extraction and annotation of liquid chromatography/mass spectrometry data sets. *Anal. Chem.* **84,** 283–289 (2012).
110. Li, C.-M. *et al.* Structural characterization of iridoid glucosides by ultra-performance liquid chromatography/electrospray ionization quadrupole time-of-flight tandem mass spectrometry. *Rapid Commun. Mass Spectrom.* **22,** 1941–1954 (2008).
111. Gupta, S. D. *Reactive Oxygen Species and Antioxidants in Higher Plants* 323 (CRC, 2010).



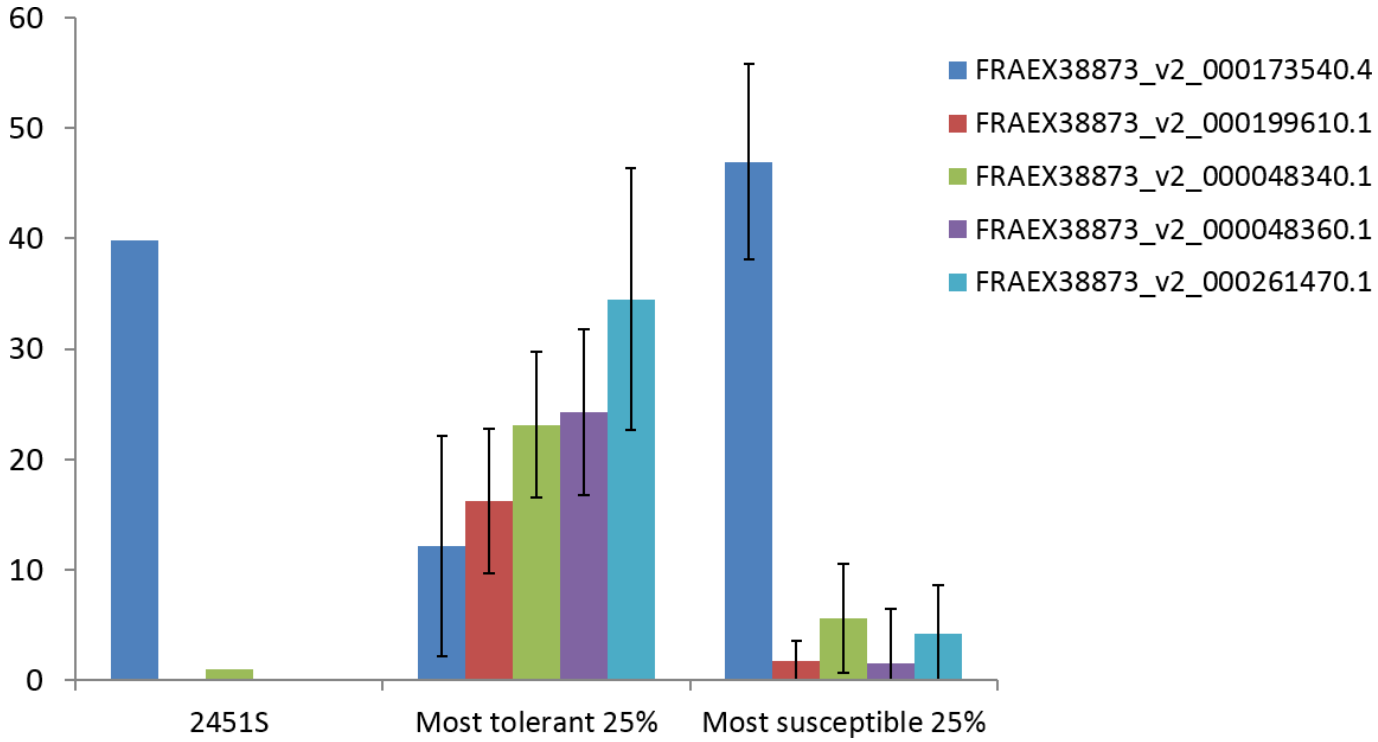
Extended Data Figure 1 | Completeness and coherence of annotation models. a, Assessment of transcript completeness for the *F. excelsior* gene set. Transcripts were classified as full-length, 5′- end, 3′-end, internal, coding (open reading frame predicted but no BLAST support), unknown (no BLAST support), mis-assembled and putative ncRNA using Full-lengtherNEXT (version 0.0.8); 76.43% of transcript models were identified as complete. **b**, Coherence in gene length between *F. excelsior* and *M. guttatus* proteins. BLAST analysis (1×10^{-5}) identified 2,576 proteins that had reciprocal best hits to 2,605 *M. guttatus* proteins identified as single copy in *M. guttatus*, *S. lycopersicum*, *S. tuberosum* and *V. vinifera* (Phytozome). A high coherence in gene length was found between *F. excelsior* and *M. guttatus*: $r > 0.917$.



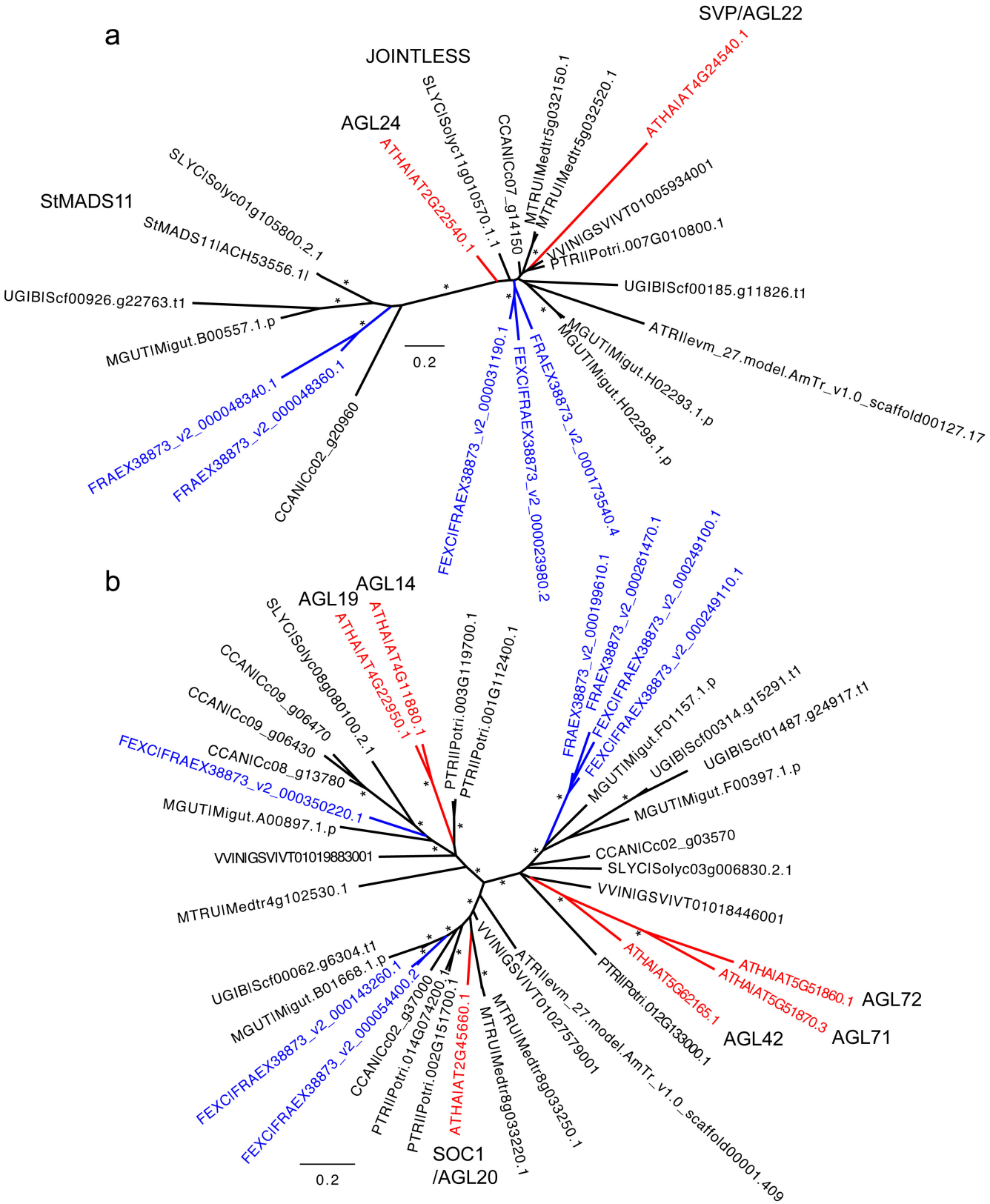
Extended Data Figure 2 | Synteny between ash and monkey flower. Syntenic dotplot between ash (vertical axis) and monkey flower (horizontal axis) showing regions of multiple synteny. Scaffolds equal to approximately 75% of the ash genome assembly for which syntenic blocks were not detected are not shown. For clarity, small scaffold names are omitted.



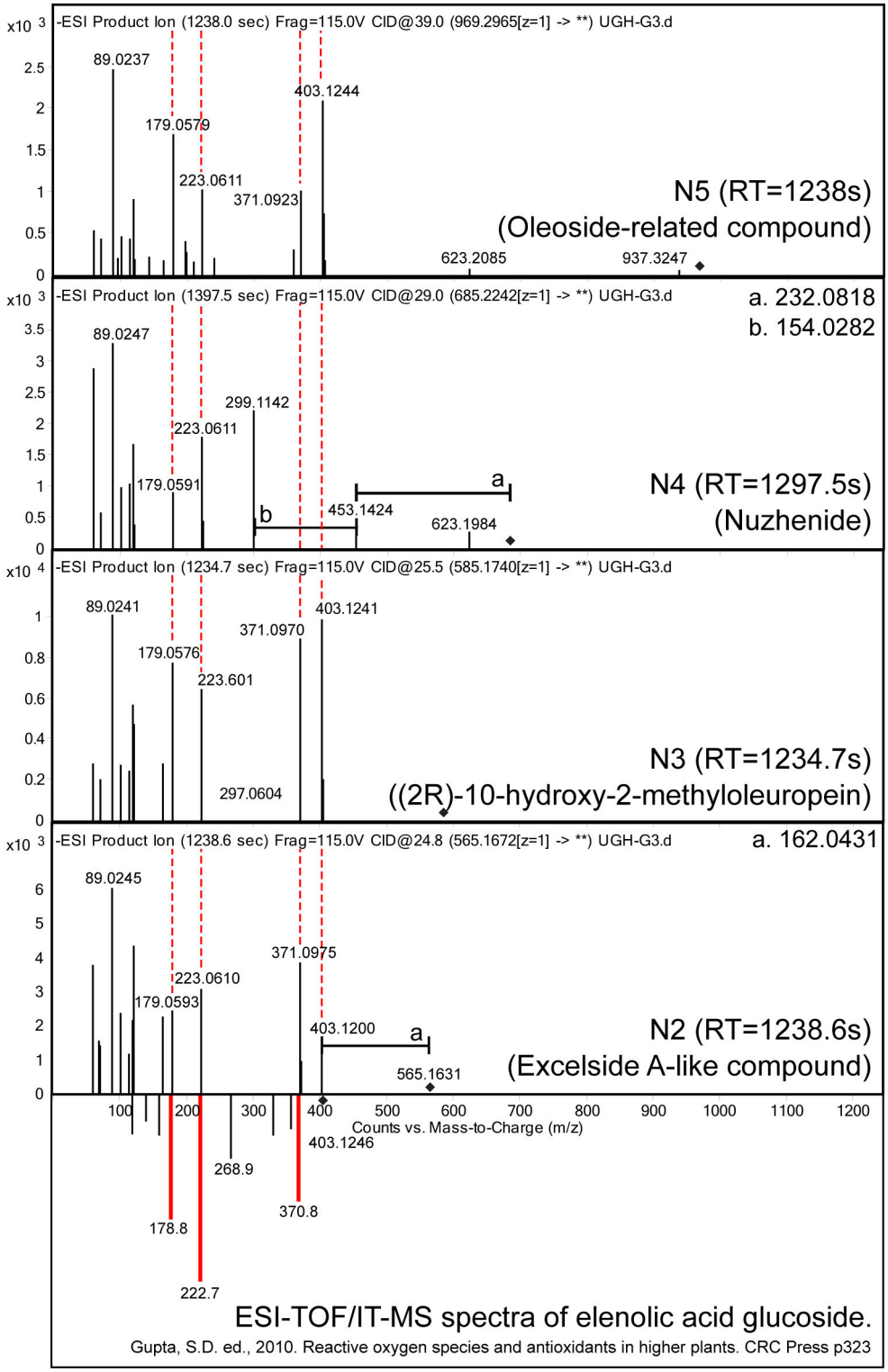
Extended Data Figure 3 | Population structure of *F. excelsior* in Europe. **a**, Results from STRUCTURE; three replicates were run for $k = 3$, with each replicate using a different set of 8,955 SNPs as input. Numbers refer to samples, whose locations are given in Supplementary Table 11. **b**, Δk values for three runs of STRUCTURE of each value of k between $k = 2$ and $k = 19$; $k = 3$ has the highest Δk value of 32.91. **c**, Effective population size history estimated using the SNeP program, with genotype information from all 38 diversity panel samples at 394,885 SNP loci.



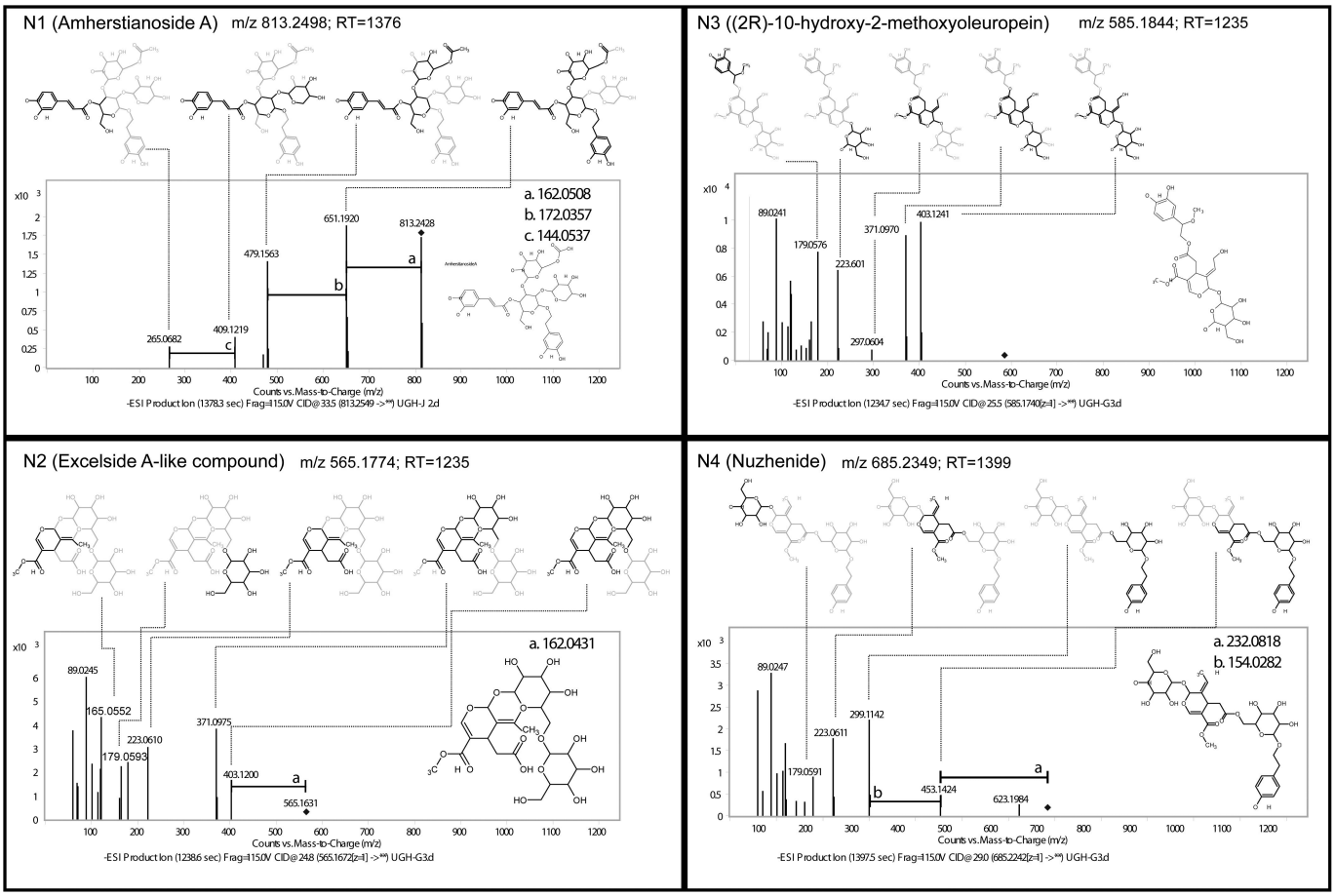
Extended Data Figure 4 | Prediction of susceptibility of reference tree. RPKM values for leaf material from the low heterozygosity reference tree 2451S for five CDS models predictive for ADB. These are shown next to expression profiles for the Danish Scoring Panel with the least susceptible and most susceptible expression patterns according to the GEM analysis.



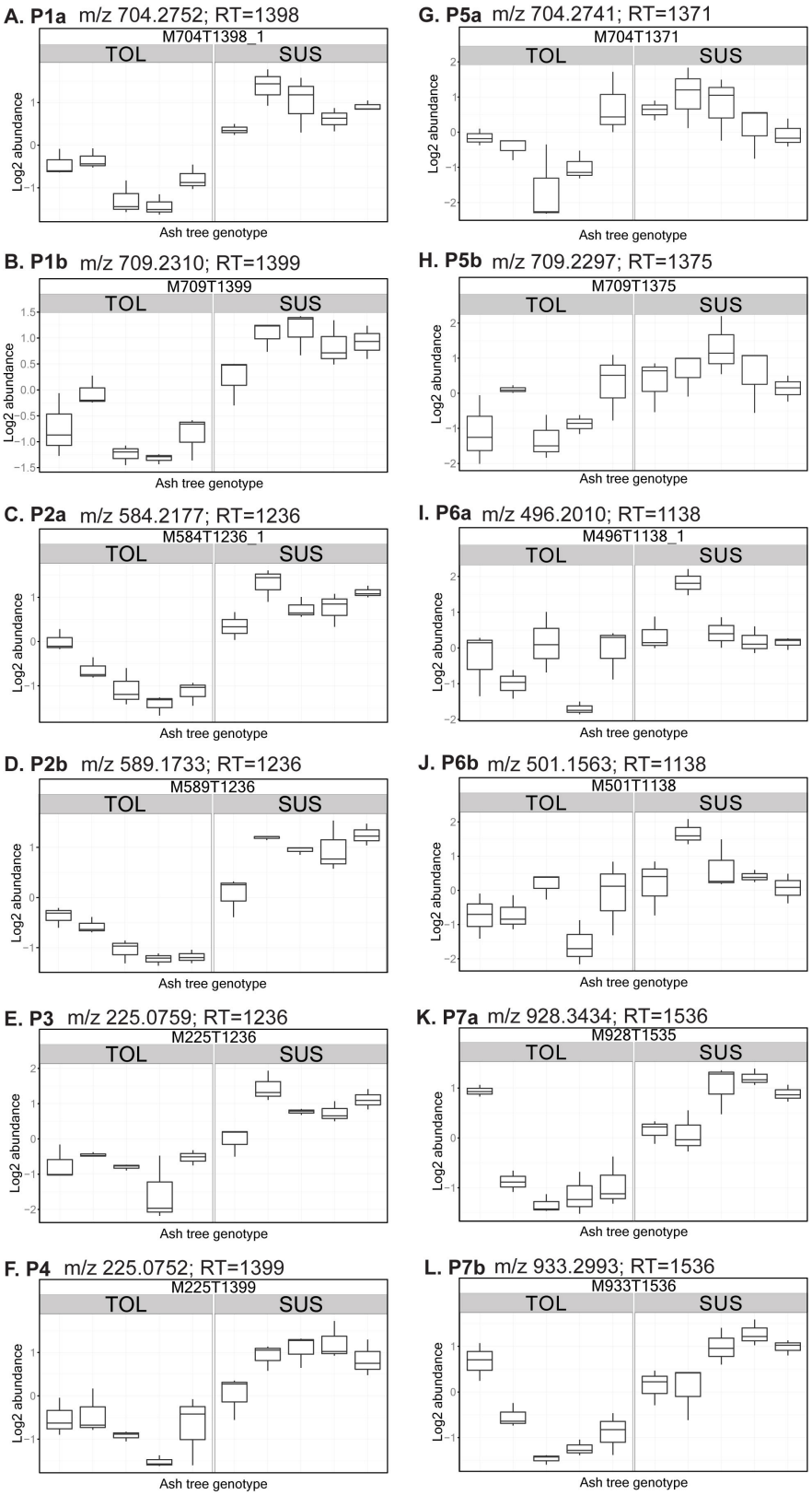
Extended Data Figure 5 | Investigation of the function of GEMs for low susceptibility to ADB. Unrooted maximum likelihood trees from the RAxML analyses. **a**, Best scoring maximum likelihood tree from the phylogenetic analysis of SVP/AGL22 and StMADS11-like sequences. **b**, Best scoring maximum likelihood tree for the SOC1-like sequences. Nodes with bootstrap support values of at least 70 from the maximum likelihood analysis and posterior probabilities of at least 0.95 from the Bayesian analysis are indicated with asterisks. *F. excelsior* sequences are shown in blue; *A. thaliana* sequences in red. Four-letter taxon codes at the start of sequence names, where present, follow those in Extended Data Table 1. Sequence names are those from the original data files used for the orthoMCL analysis (see Supplementary Table 10), with the exception of the StMADS11 protein from potato, where the GenBank accession number is given. Common names for selected genes/proteins are annotated on the trees. Scale bars indicate the mean number of substitutions per site.



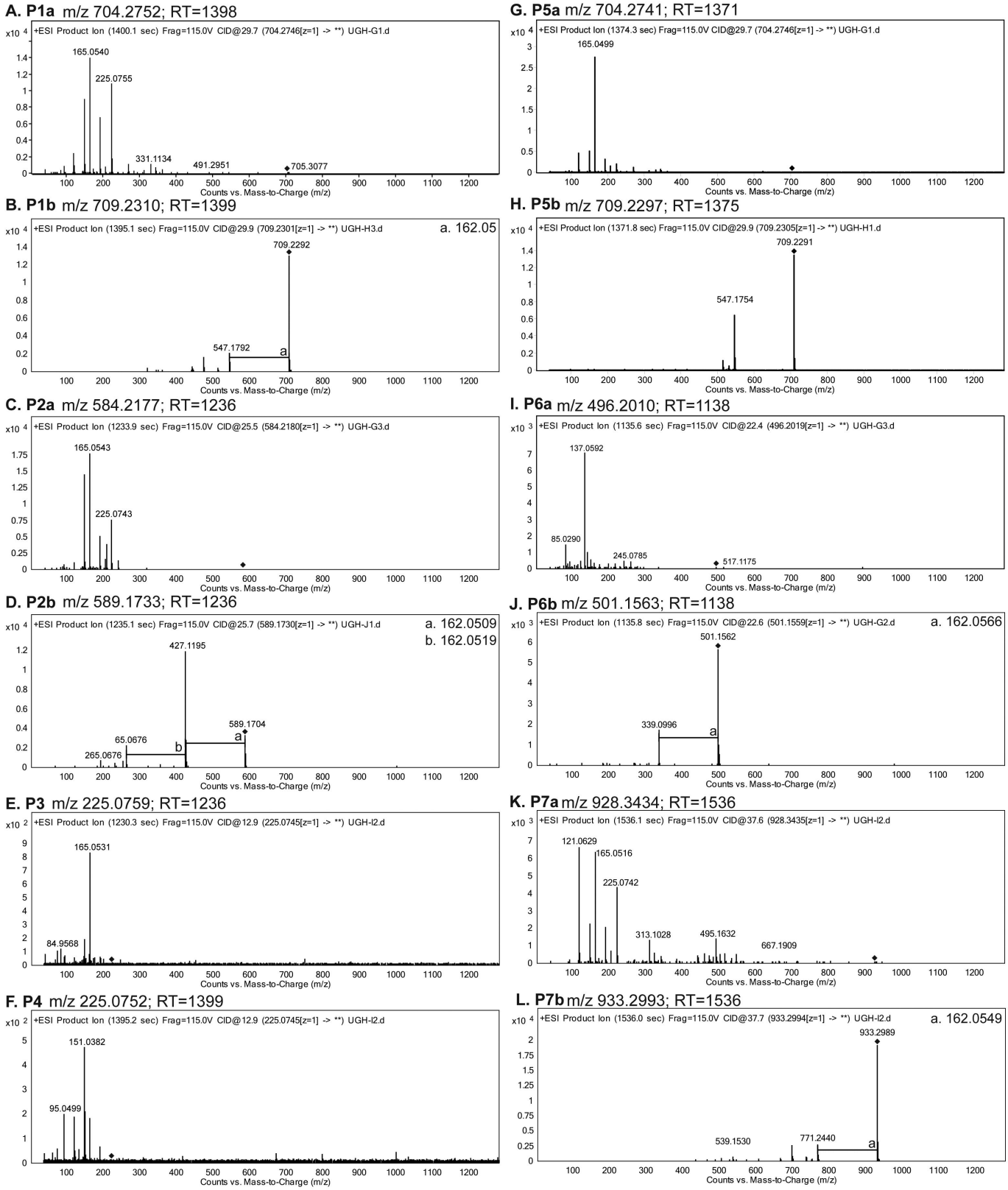
Extended Data Figure 6 | MS/MS mirror plot of elenolic acid glucoside (ESI-TOF/IT-MS) compared with four negative mode features (N_2, N_3, N_4 and N_5). The spectra share four product ions in common: m/z 179, 223, 371 and 403 (elenolic acid glucoside molecular ion). These product ions correspond to a loss of a methyl and hydroxyl group (403–371), loss of hexose (403–223), which is followed by a loss of CO_2 (223–179). Elenolic acid corresponds to the secoiridoid part of oleuropein-related compounds, suggesting that these four compounds are secoiridoids[112].



Extended Data Figure 7 | Identification of MS/MS product ions for four iridoid-glycoside-related features observed in negative mode. Predicted structure for key *m/z* peaks using Molecular Structure Correlator (Agilent) and the structure of putative identities. Bonds and atoms in black are present in that product ion, whereas grey indicates loss.



Extended Data Figure 8 | Identification of iridoid-glycoside-related metabolites in positive mode. Box-plots showing abundance (log_2 transformed) of features in positive mode discriminating between five different genotypes of high- (TOL) and low- (SUS) susceptibility ash trees.



Extended Data Figure 9 | Identification of metabolites. MS/MS fragmentation product ion data of features discriminating between five different genotypes of high- (TOL) and low- (SUS) susceptibility ash trees in positive mode. Corresponding box-plots are presented in Extended Data Fig. 8.

Extended Data Table 1 | The 20 largest clusters in *F. excelsior* from the OrthoMCL analysis of 11 species showing the number of sequences from each species belonging to the clusters

OrthoMCL cluster name	Putative gene family name(s)/ function(s)	FEXC	ATHA	ATRI	CCAN	MGUT	MTRU	PITA	PTRI	SLYC	UGIB	VVIN
OG_00001	Pentatricopeptide repeat (PPR) superfamily, Tetratricopeptide repeat (TPR)-like superfamily	102	91	35	101	103	107	212	105	93	73	118
OG_00003	Leucine-rich repeat receptor-like protein kinase family, CLAVATA1-related receptor kinase-like proteins/ protein serine/threonine kinase activity, kinase activity, ATP binding.	81	40	34	112	52	112	121	114	50	24	63
OG_00005	Subtilase family, Subtilisin-like serine endopeptidase family protein/ identical protein binding, serine-type endopeptidase activity.	63	46	42	50	95	88	40	67	71	21	65
OG_00006	**S-locus lectin protein kinase family, Putative receptor-like serine/ threonine protein kinases/ protein amino acid phosphorylation, recognition of pollen.**	**58**	**32**	**7**	**42**	**43**	**125**	**9**	**183**	**53**	**1**	**52**
OG_00007	Leucine-rich repeat protein kinase family, HIT-type Zinc finger family protein/ protein serine/threonine kinase activity, kinase activity, ATP binding protein	55	8	7	161	64	77	59	47	41	12	26
OG_00012	**Laccase family /lignin biosynthesis, cell wall biosynthesis.**	**43**	**18**	**14**	**23**	**20**	**23**	**54**	**54**	**27**	**8**	**43**
OG_00019	**Calcium dependent protein kinase family/ putative calcium sensors.**	**40**	**31**	**11**	**16**	**23**	**25**	**9**	**28**	**28**	**24**	**16**
OG_00039	**Wall-associated kinase family/ kinase activity, protein amino acid phosphorylation.**	**40**	**19**	**3**	**8**	**34**	**20**	**0**	**46**	**9**	**0**	**10**
OG_00010	**Major facilitator superfamily/ transporter activity.**	**39**	**22**	**20**	**25**	**28**	**49**	**54**	**40**	**22**	**20**	**25**
OG_00015	**P-glycoprotein family/ ATPase activity, coupled to transmembrane movement of substances.**	**37**	**22**	**14**	**25**	**23**	**39**	**38**	**36**	**22**	**10**	**20**
OG_00021	n/a	34	0	0	167	18	1	0	0	23	3	0
OG_00037	**Cellulose synthase family (CESA), Cellulose synthase-like proteins/ cell wall biosynthesis.**	**31**	**16**	**14**	**12**	**14**	**21**	**10**	**28**	**17**	**19**	**16**
OG_00004	LRR and NB-ARC, and NB-ARC domain-containing disease resistance proteins/ ATP binding, protein binding.	30	2	0	264	44	206	3	115	15	2	81
OG_00028	Cytochrome P450, family 71, subfamily B	29	28	13	35	17	47	0	26	21	0	5
OG_00026	FAD-binding Berberine family/ electron carrier activity, oxidoreductase activity, FAD binding, catalytic activity.	28	27	4	27	26	28	1	63	19	5	4
OG_00022	Putative ligand-gated ion channel subunit family/ uncharacterized functions.	28	20	24	18	37	8	24	43	11	11	21
OG_00025	Malectin/receptor-like protein kinase family, Protein kinase superfamily protein/ kinase activity, protein amino acid phosphorylation.	27	17	9	25	31	43	1	41	19	12	9
OG_00016	Pleiotropic drug resistance family, ABC-2 and Plant PDR ABC-type transporter family/ nucleoside-triphosphatase activity, ATPase activity, nucleotide binding, ATP binding.	27	16	15	23	20	33	43	29	25	13	33
OG_00059	**Leucine-rich repeat protein kinase family, Plasma membrane LRR receptor-like serine threonine kinase proteins, Somatic embryogenesis receptor-like kinase proteins/ protein serine/threonine kinase activity, kinase activity, ATP binding.**	**26**	**14**	**7**	**8**	**14**	**15**	**7**	**20**	**13**	**11**	**11**
OG_00085	**Raffinose synthase family/ carbohydrate, biosynthesis, metabolism and catabolism.**	**26**	**5**	**12**	**9**	**13**	**12**	**9**	**13**	**6**	**12**	**10**

Clusters containing at least five more sequences from *F. excelsior* than for the other asterid species (underlined) are shown in bold. FEXC, *F. excelsior*; ATHA, *A. thaliana*; ATRI, *A. trichopoda*; CCAN, *C. canephora*; MGUT, *M. guttatus*; MTRU, *Medicago truncatula*; PITA, *P. taeda*; PTRI, *P. trichocarpa*; SLYC, *S. lycopersicum*; UGIB, *U. gibba*; VVIN, *V. vinifera*. Details of gene families in column two are inferred from the gene family membership/function of *A. thaliana* genes (according to The *Arabidopsis* Information Resource; http://www.arabidopsis.org) belonging to these clusters. It should be noted that OrthoMCL clusters are not necessarily equivalent to gene families as a single gene family may be split over multiple clusters and multiple gene families may be grouped into a single cluster.

Emerging Genomics of Angiosperm Trees

Elizabeth Sollars and Richard Buggs

Abstract Genome sequence assemblies of many angiosperm trees used in forestry are now emerging, in addition to the well-characterised genomes of black poplar and eucalyptus reviewed in previous chapters of this book. Whilst the number of published genomes of angiosperm forest trees lags behind that of angiosperm trees grown commercially for fruit or nuts, many new projects are underway. This is aided by the ever-decreasing cost of DNA sequencing technologies and has diverse motivations including tree improvement, ecological and evolutionary studies. In this chapter, we briefly review a number of recent whole genome projects including Chinese chestnut, European ash, dwarf birch, pedunculate oak, purple willow and shrub willow. We also describe new projects not yet in the public domain or with non-genomic data, and list various online resources where data and information can be accessed. We discuss potential future steps in improving genome assemblies, and the uses of such information in fields such as genomic selection to assist tree breeding.

Keywords Genome sequencing • Angiosperm • Tree • Breeding • Transcriptomics

Introduction

The past decade has seen the emergence of genome projects on angiosperm forest tree species that have never before been viewed as model organisms. In many cases, the only previous genetic research upon these species was the study of population genetic diversity at a handful of loci. Unlike poplar (see Chap. 5) and eucalyptus (see Chap.6), most of these species have not yet been, and in some cases will never be, subject to intensive selection and breeding programmes. For many, their economic importance would scarcely justify a multi-million dollar genome project. But the rapid fall in the cost of sequencing since 2006 has permitted their genome sequencing at low expense, while the increasing availability of bioinformatics tools and high performance computers has made genome assembly possible even for small research groups.

E. Sollars • R. Buggs (✉)
School of Biological and Chemical Sciences, Queen Mary University of London, London, UK

Royal Botanic Gardens Kew, Richmond, Surrey, UK
e-mail: r.buggs@qmul.ac.uk

A.T. Groover and Q.C.B. Cronk (eds.), *Comparative and Evolutionary Genomics of Angiosperm Trees*, Plant Genetics and Genomics: Crops and Models,
DOI 10.1007/7397_2016_16

Biologists with specific interests from evolution to epidemiology have been able to sequence the genomes of their study species as foundational data for their research.

Such projects bring with them prospects of using genetic markers ascertained from a reference genome sequence to breed improved genotypes. Whether it be for desirable traits such as yield and wood quality, or defensive traits such as disease resistance, utilising genetic information will speed up breeding in tree species with typical generation times of above 10 years (Neale and Kremer 2011). Even when breeding is not in view, genomic data can greatly enhance our understanding of local adaptation and ecological genetics in native tree species (Plomion et al. 2016; Neale and Kremer 2011).

In this chapter, we review the emerging genome sequences for six tree species of which we are aware to have genome assemblies (excluding fruit and nut trees): Chinese chestnut (*Castanea mollissima*), pedunculate oak (*Quercus robur*), European ash (*Fraxinus excelsior*), purple willow (*Salix purpurea*), shrub willow (*Salix suchowensis*), and dwarf birch (*Betula nana*) (See Table 1). Several of these have been placed in the public domain before publication of an associated paper, to the benefit of the research community (Neale et al. 2013). We mention other genome sequencing projects of which we are aware, but are not yet in the public domain. We also highlight other species for which transcriptomes, genetic maps or genome-wide marker data are available, and which are likely candidates for future whole genome sequencing projects. We mention only in passing the numerous genome projects on fruit and nut trees, as the main foci of these projects are agronomic rather than forestry or wood product-related. Inevitably, there is likely to be interesting research on some forest tree species that we are unaware of, and which will therefore be missed in this review.

Chinese Chestnut (*Castanea mollissima*)

The Chinese chestnut has received particular attention as a genomic resource because the species is resistant to chestnut blight, a disease caused by the pathogenic fungus *Cryphonectria parasitica (*Anagnostakis 1987*)*. This fungus has devastated American chestnuts, which are highly susceptible, since its introduction to the USA around 1904 (Anagnostakis 1987). Considerable effort has gone into breeding American chestnut trees with resistance to the fungus either through hybridising American with Japanese or Chinese chestnuts, or by using transgenics to introduce resistance genes into the American chestnut genome (Hebard et al. 2014).

Genetic and physical maps of the *Castanea mollissima* (~800 Mbp) genome were published in 2013 (Kubisiak et al. 2013; Fang et al. 2013). A consortium led by John Carlson at Penn State University has assembled a genome sequence of *C. mollissima* using a combination of 454 and Illumina MiSeq reads, and BAC paired-end Sanger sequences (Carlson 2014), with scaffolds anchored into pseudo-chromosomes using the physical map. They also annotated the genome with over 36,000 gene models, which are available on the Hardwood Genomics webpage (www.hardwoodgenomics.org). The consortium has sequenced additional genotypes of Chinese and American chestnut to obtain variant data (Carlson 2014). A research group based at Purdue University has resequenced 16 Chinese chestnuts and hybrids

with variable blight resistance, in order to investigate variation at loci implicated in resistance (LaBonte and Woeste 2016). In addition, transcriptomic data are available for American chestnut (*C. dentata*), Japanese chestnut (*C. crenata*), and European chestnut (*C. sativa*), generated under The Fagaceae Project in the USA (http://www.fagaceae.org/).

Pedunculate Oak (*Quercus robur*)

Quercus robur is one of Europe's most widespread trees, with considerable timber and ecological value. Due to its long generation time, the focus of genomic research on this species has been upon ecological and speciation genetics, rather than selection and breeding for trait improvement (Plomion et al. 2016). A genome sequence, based on Sanger, 454, and Illumina read data, is being assembled by a French collaboration led by Christophe Plomion at INRA (the French National Institute for Agricultural Research) in Bordeaux (Plomion et al. 2016). The 1C-genome size of pedunculate oak is ~740 Mbp (Lesur et al. 2011). A tree was chosen for sequencing for which genetic maps (Durand et al. 2010; Bodénès et al. 2012) and a small amount of exploratory genomic sequence data (Lesur et al. 2011; Faivre-Rampant et al. 2011) had already been generated (Lesur et al. 2011; Faivre-Rampant et al. 2011). This tree had been used as a parent in crosses to study quantitative traits (Plomion et al. 2016). An initial assembly was released in November 2014 on the project's website, http://www.oakgenome.fr, consisting of 17,910 scaffolds (Plomion et al. 2016). The second, improved version consists of approximately 1400 scaffolds ordered into 12 pseudomolecules with the help of a linkage map (C. Plomion, pers. comm.). The consortium has also generated genome sequence data for four other European oak species (Plomion et al. 2016). In addition, transcriptomic data are available for red oak (*Quercus rubra*) and white oak (*Quercus alba*) generated under The Fagaceae Project in the USA (http://www.fagaceae.org/).

European Ash (*Fraxinus excelsior*)

The European ash, *Fraxinus excelsior*, is a widespread woodland tree in Europe of ecological and economic consequence. Though it had previously been subject to little molecular genetic research, apart from investigations of population structure and mating systems, it became the subject of genome sequencing in 2012 due to the rapid spread of the fungal pathogen *Hymenoscyphus fraxineus* across Europe (Pautasso et al. 2013). The ~880 Mbp genome of a low heterozygosity British tree produced by self-pollination was sequenced using 454 and Illumina technologies by a collaboration between Queen Mary University of London and Eurofins MWG GmbH with assemblies released at www.ashgenome.org. Meanwhile, a Danish tree that had been shown to have low susceptibility to *H. fraxineus* (McKinney et al. 2011) was Illumina sequenced at The Genome Analysis Centre, Norwich and released at https://geefu.

oadb.tsl.ac.uk/. Collaboration between the two institutions led to the annotation of the reference genome from the low heterozygosity tree, which proved easier to assemble, and the low coverage sequencing of 37 further *F. excelsior* trees representing provenances from across Europe to study natural variation (E. Sollars et al., in press). The reference genome has facilitated associative transcriptomic studies, identifying gene expression markers associated with reduced susceptibility to *H. fraxineus* in Denmark (Harper et al. 2016). In the USA, where ash populations are being devastated by the emerald ash borer, transcriptome sequencing has been conducted on green ash (*F. pennsylvanica*) and white ash (*F. americana*) within the Hardwood Genomics Project (http://hardwoodgenomics.org/). Low coverage genome sequencing of green ash has also been generated under this project.

Purple Willow (*Salix purpurea*)

Salix purpurea has become a key model species in genetic improvement for shrub willows, for use as a biomass crop. A genome project is led by Larry Smart's group at Cornell University and involves researchers from Oak Ridge National Laboratory and the J. Craig Venter Genome Institute. The ~450 Mbp genome has been sequenced using the Illumina platform and assembled into scaffolds which have been annotated using RNA-seq data and anchored to a genetic map (Carlson et al. 2014). This has been released at http://phytozome.jgi.doe.gov/. The genome has already been used as a reference for genotyping by sequencing (GBS) of 100 of further individuals, leading to the identification of genetic markers associated with growth and biomass yield (Carlson et al. 2016; Gouker et al. 2016).

Shrub Willow (*Salix suchowensis*)

The genome of *Salix suchowensis* was published in 2014 by research institutes in China, USA, and the UK (Dai et al. 2014). It consists of ~304 Mbp of sequence in 103,144 scaffolds, on which 26,599 putative protein-coding genes were annotated. They also compare the genome to that of poplar (*Populus trichocarpa*), and investigate divergence, substitution rates, and whole genome duplications in the two species. Since the publication of the genome sequence, it has been used in a genome-wide study of heat shock proteins (Zhang et al. 2015). The researchers identified 27 HSPs and studied their expression profiles during development and in response to abiotic and biotic stresses such as heat, drought, and salt.

Dwarf Birch (*Betula nana*)

Dwarf birch is a small tree found in boreal scrub communities; one of the most northerly distributed woody angiosperms. Though of little economic importance, it is a keystone species to the ecology of the sub-arctic. It also holds promise as a

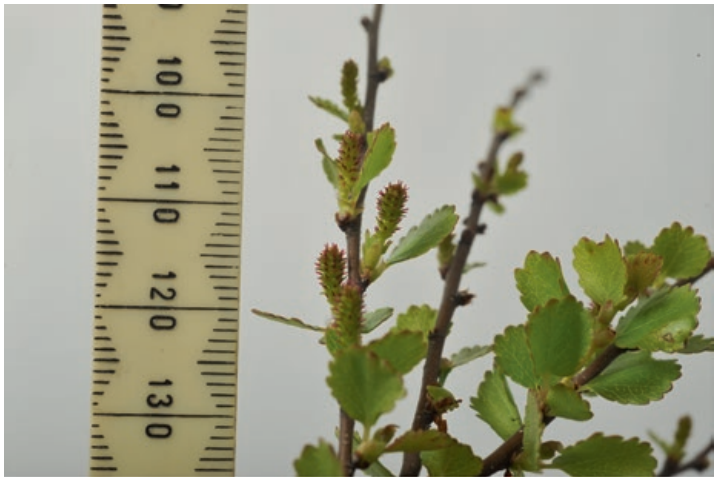


Fig. 1 Sequenced individual of dwarf birch (*Betula nana*), a species found in boreal regions. The small size of dwarf birch allows it to be cultivated in limited space. Scale in mm

Table 1 Assembly statistics for six selected emerging angiosperm tree genomes

Species	1C genome size (mbp)	Assembly version	No. scaffolds	Assembly size	Scaffold N50 (kbp)	No. genes annotated
C. mollissima	794	v1.1	41,260	724 mbp	39.6	36,478
Q. robur	740	v1.0	18,000 > 2000 bp	1350 mbp	257	54,000
F. excelsior	877	v0.5	89,487 nuclear ≥ 500 bp 26 mitochondrial 1 plastid	868 mbp	104.0	38,852
S. purpurea	450	v1.0	7528	392 mbp	17,359.0	37,865
S. suchowensis	429		103,144 ≥ 100 bp 7516 ≥ 2000 bp	304 mbp	925	26,599
B. nana	450		551,923 75,763 ≥ 500 bp	564 mbp	18.79	None

model organism, being small in size and short in generation time, with a ~450 Mbp 1C-genome size. A draft genome was published in 2013 (Wang et al. 2013), of the individual shown in Fig. 1, based on Illumina sequencing. Though fragmented and preliminary, this assembly was a useful reference for the restriction amplified digest (RAD) sequencing of other individuals of *B. nana*, and also *B. pubescens* and *B. pendula* (Wang et al. 2013). An improved assembly using SMRT sequencing (Pacific Biosciences, CA, USA) is underway).

Genomes of Angiosperm Fruit and Nut Trees

Whilst this chapter is mainly focused on angiosperm trees related to forestry or biomass production, it must be noted that a wealth of genomic data is being generated on other angiosperm trees used for agronomic purposes. Genome sequences have been assembled for species such as: walnut (*Juglans regia*) (Martínez-García et al. 2015), European

hazelnut (*Corylus avellana*) (Rowley et al. 2012; Rowley 2016), several citrus fruit (*Citrus*) genomes (Wu et al. 2014; Xu et al. 2013), apple (*Malus domestica*) (Velasco et al. 2010), peach (*Prunus persica*) (Verde et al. 2013), Chinese white pear (*Pyrus* x *bretschneideri*) (Wu et al. 2013), European pear (*Pyrus communis*) (Chagné et al. 2014), pistachio (*Pistachia vera*) (Kafkas 2016), cacao (*Theobroma cacao*) (Argout et al. 2011), coffee (*Coffea canephora*) (Denoeud et al. 2014), papaya (*Carica papaya*) (Ming et al. 2008), date palm (*Phoenix dactylifera*) (Al-Dous et al. 2011; Al-Mssallem et al. 2013; Mathew et al. 2014), and oil palm (*Elaeis guineensis*) (Singh et al. 2013). The genome of the rubber tree (*Hevea brasiliensis*), used for its latex production, has also been assembled and published (Rahman et al. 2013). In addition there are several projects assembling the olive (*Olea europaea*) genome; the International Olive Genome Consortium (http://olivegenome.karatekin.edu.tr/), the OLEA consortium (http://www.oleagenome.org/), and the Olive Tree Genome Project (http://olive.crg.eu, Cruz et al. 2016).

These datasets are important in informing comparative studies of tree genomes generally and in some cases provide useful reference genomes for timber trees, such as walnut cultivars grown for forestry rather than nut production, and rubber trees used for timber once latex productivity has declined.

Projects Not Yet in the Public Domain

There are also angiosperm tree genome projects that we know to be underway but which have at the time of writing not yet released data into the public domain. By their nature, such projects are little publicised. The authors are aware of a project in Finland on *Betula pendula* (silver birch) (Rajaraman and Salojärvi 2015) and another in China on *Betula platyphylla* (Japanese white birch) (C. Yang pers. Comm., see http://birch.genomics.cn/). There are no doubt other projects on genera that we do not work on and thus are unaware of.

Species with Genome-Wide Data

In addition to the whole genome sequence assemblies outlined above, genome-wide data is now available for many other trees. The Dendrome (http://dendrome.ucdavis.edu/) portal provides comprehensive and continually updated access to these rapidly growing resources, many of which have not yet been accompanied by a published paper (Wegrzyn et al. 2008). In several cases, these datasets have been collected to accompany and aid the interpretation of the reference genomes we have already described, such as the population datasets mentioned above for oak species in Europe, European ash, and American chestnut and hybrids. Here, we will not catalogue everything that is available, but mention a selection of tree species which are the first in their genus to be subject to sequencing, and may therefore emerge as reference sequences for their genus as and when funding becomes available. The USA-based Fagaceae Project has released 454 transcriptomic data for American beech (*Fagus grandifolia*) (http://www.fagaceae.org/), as well as generating data for oaks and chestnuts. European

groups have also released transcriptomic data for *Fagus sylvatica* on Dendrome. Low coverage genome sequencing was recently reported for ten native hardwood tree species from the eastern United States including: blackgum (*Nyssa sylvatica*), redbay (*Persea borbonia*), sugar maple (*Acer saccharum*), sweetgum (*Liquidambar styraciflua*), and honeylocust (*Gleditisia triacanthos*) (Staton et al. 2015).

Future Steps

Rapid progress has been made in angiosperm forest tree genomes over the past few years due to the rise of next generation sequencing, and the pace of progress is set only to increase as new technologies such as Oxford Nanopore (Oxford Nanopore Technologies, Oxford, UK) sequencing, SMRT sequencing (Pacific Biosciences, CA, USA), and optical mapping continue to improve (Howe and Wood 2015; VanBuren et al. 2015). This raises the question as to where researchers need to focus their future efforts. To some extent, research programmes on poplar (see Chap. 5) and eucalyptus (see Chap. 6) provide good exemplars, though less funding may be available for emerging genomes than was available for those species. Several possibilities are available to researchers, which we outline below.

Firstly, improvement of current reference genome assemblies is needed. Most of the genome sequences reviewed above are still in fragmented states and far from being assembled at a chromosomal level. Some are lacking genetic maps to anchor scaffolds to, and where maps are available, a high proportion of scaffolds remain unanchored. Furthermore it is well known that *de novo* genome assemblies often miss genes, spuriously duplicate them, or join contigs erroneously (Denton et al. 2014; Elsik et al. 2014; Alkan et al 2011). Difficult decisions need to be taken regarding how much time and effort to devote to improving reference genomes. A better genome assembly may lead to more powerful genome-wide analyses of, for example, trait-associated loci or patterns of introgression. However, in some cases, especially where the reference tree selected is highly heterozygous, the genome contains many repetitive elements, or is polyploid, a chromosomal-level genome assembly may be almost impossible with current technologies. In these cases, and perhaps more widely, the wisest course may be to do the best possible assembly with 200× Illumina coverage and as much longer read data that can be afforded (such as 454 or PacBio reads), and then wait for new technologies to develop or improve (such as Oxford Nanopore technology at the time of writing (Goodwin et al. 2015)) before attempting to make substantial improvements to the assembly. Enhancements of sufficient magnitude are likely to stem from technologies that focus on joining and ordering scaffolds, such as optical mapping and BioNano Irys, or filling in assembly gaps that the usual technologies are unable to sequence, rather than from simply obtaining additional sequencing data.

Secondly, researchers can focus on sequencing more individuals from the same species. This has been done in many of the species mentioned above, such as ash, chestnut, oak, and willow, where a focal sequence now has a penumbra of additional genomes, often sequenced at lower coverage and assembled using the focal individual

as a reference. There is a danger that differences in gene content among individuals, where such variation exists, may be undiscovered by this approach, but it does allow the characterisation of much genome-wide variation within species, which may aid the development of SNP panels for larger studies of population variation. Without population-level studies, bottom-up inferences of the functional significance of loci within the genome are impossible, and all we can rely on are homology searches to better functionally-characterised plant genomes. These are of course useful for initial annotation of plant genomes, but rely heavily on the assumption that similar sequences will have similar functions. Such approaches are of limited value for the characterisation of taxonomically restricted genes (Khalturin et al. 2009).

Thirdly, sequencing of other species of the same genus can be undertaken, as this may allow the characterisation of genes responsible for key differences between species, such as ecological adaptations in *Q. robur* versus *Q. petraea* or blight resistance in *C. mollissima* versus *C. dentata* (see above). Here, the independent assembly of the different genomes may be particularly important, in order to take account of species-specific genes and gene-family expansions. Reference-guided assembly approaches have been developed which assemble genome sequences independently of a related reference so as to retain any variation, but still use the reference to aid the placing of scaffolds into longer contiguous sequences (Bao et al 2014; Kim et al. 2013). On the other hand, mapping of reads from closely related species to a single reference genome can identify population dynamics such as hybridisation and introgression (Suarez-Gonzalez et al. 2016).

Fourthly, research could focus on functional characterisation of genes within the reference genome using experimental approaches. However, such approaches, which have worked well for *Arabidopsis*, are highly challenging in tree species. Generation of inbred lines, knock-out lines, or multiple mapping populations are seldom feasible for long-lived trees which outgrow laboratory growth cabinets long before reproduction is a possibility. In some cases, experiments can be carried out using orthologues in a model plant species (Salmon et al. 2014). However, newly developed, targeted methods could easily achieve what would take years with conventional approaches. For example, targeted mutagenesis by CRISPR/Cas9 has been used to create knockout mutations in *Populus tomentosa* (Chinese white poplar) (Fan et al. 2015), and the expression of genes was inhibited using virus-induced gene silencing in two other *Populus* species (Shen et al. 2015).

A related consideration is at what point in time research groups should publish their genome assemblies. In recent years there has been a growing willingness to release data early (Neale et al. 2013), but it is notable that four of the six major reference genomes reviewed above do not have a final peer-reviewed publication associated with them, even though some of the assemblies have now been available online for years. This may be due to a lack of funding meaning that personnel are not available for the higher-level analyses and manuscript writing that is needed. It may be due to consortia waiting for further improvements to the assembly and annotation before they attempt a high-impact publication: the presubmission paper for *Q. robur* (Plomion et al. 2016) outlines such a strategy. It may be due to lengthy review processes by journals. Indeed, the only genomes of those reviewed here with a final publication at present are *B. nana* (which follows the opposite extreme of publishing a very preliminary and highly

fragmented genome with no annotation (Wang et al. 2013)) and *S. suchowensis* (Dai et al. 2014). It would seem that there are currently as many different publishing strategies for tree genomes as there are reference assemblies.

Decisions about when to stop improving a genome, or when to publish can be informed by quality metrics. Various statistics are often used to compare assemblies and allow optimisation of a series of variable parameters used during assembly, or to look for notable improvement upon receiving additional data. Simple measurement statistics such as the length of assembly, number of scaffolds, and N50 give an indication of contiguity, but cannot describe the quality of the sequence. Computational methods that search the assembly for conserved genes (Simão et al. 2015; Parra et al 2007), or use the mapping of DNA reads can be used for this purpose (Vezzi et al 2012; Clark et al. 2013). However, given that some genomes are harder to assemble than others, and that different uses of genome assemblies require different levels of quality, there is no single standard for when a genome assembly is considered publishable.

Broader Implications

Angiosperm forest trees in the past may have appeared to be less promising species for genetic improvement than their gymnosperm counterparts. However, emerging genomes show that there is an important place for genomic research on angiosperm forest trees. Because their genome sizes are much smaller than the multi-gigabase genomes of conifers (Kelly et al. 2012), they are much quicker and cheaper to sequence. The sequencing of multiple individuals is also feasible. Genomic research is thus more likely to yield rapid benefits for angiosperm trees than for gymnosperms.

Herbaceous angiosperms have tended to dominate genomic research, as they are far more amenable to experimentation and breeding than angiosperm trees. Recent years have seen the successful application of genomic selection to annual crop species (Heffner et al 2009), which has allowed time and money to be saved in breeding due to selection of seedlings at early, pre-reproductive ages. If such approaches are economically viable because they can save a few weeks per generation in annual crop breeding, they may have a huge impact on tree selection and breeding (Denis and Bouvet 2012; Resende et al. 2012), where they can allow selection to take place many years before a tree is of reproductive age. Therefore, genomic selection may ultimately have a bigger impact on angiosperm tree breeding than on angiosperm crops.

As well as holding great promise, genomic research on a wide range of angiosperm trees is necessary and timely (Neale and Kremer 2011). Global trade and lax biosecurity measures have resulted in unprecedented spread of pests and diseases around the globe in the past decades, causing grave damage to tree populations (Brasier 2008; Boyd et al. 2013). Moreover, the value of natural capital and the need for renewable energy sources and carbon fixation is now appreciated more than ever (Helm 2015). It has recently been suggested that the replacement of angiosperm woodland with coniferous plantations may have contributed to, rather than mitigating, climate change (Naudts et al. 2016). As an emerging field, the genomics of angiosperm trees has much to offer our planet (Table 2).

Table 2 A selection of emerging tree genome projects

Tree	Species	Lead Researcher(s)/group	Data available	URL
Ash	*Fraxinus excelsior*	The British Ash Tree Genome Project, Richard Buggs, QMUL	WGS, genome assembly, SSRs, RNA-seq, gene annotation, bisulphite-seq (ongoing)	www.ashgenome.org
		The Nornex Consortium, Allan Downie, John Innes Centre	WGS, genome assembly, RNA-seq, gene annotation	https://geefu.oadb.tsl.ac.uk
	Fraxinus pennsylvanica	Hardwood Genomics Project	WGS, SSRs, RNA-seq	www.hardwoodgenomics.org
	Fraxinus americana		WGS, SSRs	
Beech	*Fagus grandifolia*	The Fagaceae Genomics Project	RNA-seq, EST assembly	www.fagaceae.org
Birch	*Betula nana*	Richard Buggs, QMUL	WGS, genome assembly, RAD-seq	www.birchgenome.org
	Betula platyphylla	Chunping Yang, Northeastern Forestry University, China	WGS, genome assembly, gene annotation	http://birch.genomics.cn
Black Cherry	*Prunus serotina*	Hardwood Genomics Project	WGS, SSRs	www.hardwoodgenomics.org
Black Walnut	*Juglans nigra*	Hardwood Genomics Project	WGS, SSRs, RNA-seq, ddRADtag (ongoing)	www.hardwoodgenomics.org
Blackgum	*Nyssa sylvatica*	Hardwood Genomics Project	WGS, SSRS, RNA-seq	www.hardwoodgenomics.org
Chestnut	*Castanea crenata*	The Fagaceae Genomics Project	EST assembly	www.fagaceae.org
	Castanea dentata		EST assembly	
	Castanea mollissima		WGS, EST assembly, Physical map	
	Castanea sativa		EST assembly	

Honeylocust	*Gleditsia triacanthos*	Hardwood Genomics Project	WGS, SSRs, RNA-seq,GBS (ongoing)	www.hardwoodgenomics.org
Oak	*Quercus alba*	Hardwood Genomics Project	WGS, SSRs, EST assembly	www.hardwoodgenomics.org
	Quercus robur	Christophe Plomion, INRA	WGS, genome assembly, SNPs, transcriptome assembly, genetic map	https://w3.pierroton.inra.fr/QuercusPortal/index.php www.oakgenome.fr
	Quercus rubra	Hardwood Genomics Project	WGS, RNA-seq, ddRADtag (ongoing)	www.hardwoodgenomics.org
Redbay	*Persea borbonia*	Hardwood Genomics Project	WGS, SSRs	www.hardwoodgenomics.org
Sugar Maple	*Acer saccharum*	Hardwood Genomics Project	WGS, SSRs, RNA-seq	www.hardwoodgenomics.org
Sweetgum	*Liquidambar styraciflua*	Hardwood Genomics Project	WGS, SSRs, RNA-seq	www.hardwoodgenomics.org
Tulip poplar	*Liriodendron tulipifera*	Hardwood Genomics Project	RNA-seq, GBS (ongoing)	www.hardwoodgenomics.org
Willow	*Salix purpurea*	Larry Smart, Cornell University	WGS, genome assembly, gene annotation	https://phytozome.jgi.doe.gov/pz/portal.html
	Salix suchowensis	Tongming Yin, Nanjing Forestry University	WGS, genome assembly, EST assembly	http://www.ncbi.nlm.nih.gov/bioproject/203514

References

Al-Dous EK, George B, Al-Mahmoud ME, Al-Jaber MY, Wang H, Salameh YM, Al-Azwani EK, et al. De novo genome sequencing and comparative genomics of date palm (*Phoenix dactylifera*). Nat Biotechnol. 2011;29(6):521–7.

Alkan C, Sajjadian S, Eichler EE. Limitations of next-generation genome sequence assembly. Nat Methods. 2011;8(1):61–5.

Al-Mssallem IS, Hu S, Zhang X, Lin Q, Liu W, Tan J, Yu X, et al. Genome sequence of the date palm *Phoenix Dactylifera* L. Nat Commun. 2013;4:2274.

Anagnostakis SL. Chestnut blight: the classical problem of an introduced pathogen. Mycologia. 1987;79(1):23–37. Mycological Society of America: 23–37.

Argout X, Salse J, Aury J, Guiltinan MJ, Droc G, Gouzy J, Allegre M, et al. The genome of *Theobroma cacao*. Nat Genet. 2011;43(2):101–8.

Bao E, Jiang T, Girke T. AlignGraph: algorithm for secondary de novo genome assembly guided by closely related references. Bioinformatics. 2014;30(12):i319–28.

Bodénès C, Chancerel E, Gailing O, Vendramin GG, Bagnoli F, Durand J, Goicoechea PG, et al. Comparative mapping in the Fagaceae and beyond with EST-SSRs. BMC Plant Biol. 2012;12:153.

Boyd IL, Freer-Smith PH, Gilligan CA, Godfray HCJ. The consequence of tree pests and diseases for ecosystem services. Science. 2013;342(6160):1235773.

Brasier CM. The biosecurity threat to the UK and global environment from international trade in plants. Plant Pathol. 2008;57(5):792–808. Blackwell Publishing Ltd: 792–808.

Carlson JE. The chestnut genome project. In: Plant and animal genome XXII conference. Plant and Animal Genome. 2014. https://pag.confex.com/pag/xxii/webprogram/Paper9777.html.

Carlson CH, Gouker FE, Serapiglia MJ, Tang H, Krishnakumar V, Town CD, Tuskan GA, et al. Annotation of the *Salix purpurea* L. genome and gene families important for biomass production. In: Plant and animal genome XXII conference. Plant and Animal Genome. 2014. https://pag.confex.com/pag/xxii/webprogram/Paper12085.html.

Carlson CH, Gouker FE, DiFazio S, Zhou R, Smart L. High-resolution mapping of biomass-related traits in shrub willow (*Salix purpurea* L.). In: Plant and animal genome XXIV conference. Plant and Animal Genome. 2016. https://pag.confex.com/pag/xxiv/webprogram/Paper21612.html.

Chagné D, Crowhurst RN, Pindo M, Thrimawithana A, Deng C, Ireland H, Fiers M, et al. The draft genome sequence of European pear (*Pyrus communis* L. 'Bartlett'). PLoS One. 2014;9(4):e92644.

Clark SC, Egan R, Frazier PI, Wang Z. ALE: a generic assembly likelihood evaluation framework for assessing the accuracy of genome and metagenome essemblies. Bioinformatics. 2013;29(4):435–43.

Cruz F, Julca I, Gómez-Garrido J, Loska D, Marcet-Houben M, Cano E, Galán B, et al. Genome sequence of the olive tree, Olea europaea. GigaScience. 2016;5:29.

Dai X, Hu Q, Cai Q, Feng K, Ye N, Tuskan GA, Milne R, et al. The willow genome and divergent evolution from poplar after the common genome duplication. Cell Res. 2014;24(10):1274–7.

Denis M, Bouvet J-M. Efficiency of genomic selection with models including dominance effect in the context of *Eucalyptus* breeding. Tree Genet Genomes. 2012;9(1):37–51. Springer-Verlag: 37–51.

Denoeud F, Carretero-Paulet L, Dereeper A, Droc G, Guyot R, Pietrella M, Zheng C, et al. The coffee genome provides insight into the convergent evolution of caffeine biosynthesis. Science. 2014;345(6201):1181–4.

Denton JF, Lugo-Martinez J, Tucker AE, Schrider DR, Warren WC, Hahn MW. Extensive error in the number of genes inferred from draft genome assemblies. PLoS Comput Biol. 2014;10(12):e1003998.

Durand J, Bodénès C, Chancerel E, Frigerio J-M, Vendramin G, Sebastiani F, Buonamici A, et al. A fast and cost-effective approach to develop and map EST-SSR markers: oak as a case study. BMC Genomics. 2010;11:570.

Elsik CG, Worley KC, Bennett AK, Beye M, Camara F, Childers CP, de Graaf DC, et al. Finding the missing honey bee genes: lessons learned from a genome upgrade. BMC Genomics. 2014;15:86.

Faivre-Rampant P, Lesur I, Boussardon C, Bitton F, Bodénès C, Le Provost G, Bergès H, Fluch S, Kremer A, Plomion C. Analysis of BAC end sequences in oak, providing insights into the composition of the genome of this keystone species. BMC Genomics. 2011;12:292.

Fan D, Liu T, Li C, Jiao B, Li S, Hou Y, Luo K. Efficient CRISPR/Cas9-mediated targeted mutagenesis in *Populus* in the first generation. Sci Rep. 2015;5:12217.

Fang G-C, Blackmon BP, Staton ME, Nelson CD, Kubisiak TL, Olukolu BA, Henry D, et al. A physical map of the Chinese chestnut (*Castanea mollissima*) genome and its integration with the genetic map. Tree Genet Genomes. 2013;9(2):525–37. Springer: 525–37.

Goodwin S, Gurtowski J, Ethe-Sayers S, Deshpande P. Oxford Nanopore sequencing and de novo assembly of a eukaryotic genome. BioRxiv. 2015. www.biorxiv.org, http://biorxiv.org/content/early/2015/01/06/013490.short.

Gouker FE, Zhou R, Evans L, DiFazio S, Bubner B, Zander M, Smart L. Genotypic-phenotypic variation and marker-based heritability estimates of a shrub willow (*Salix purpurea*) association population. In: Plant and animal genome XXIV conference. Plant and Animal Genome. 2016. https://pag.confex.com/pag/xxiv/webprogram/Paper19730.html.

Harper AL, McKinney LV, Nielsen LR, Havlickova L, Li Y, Trick M, Fraser F, et al. Molecular markers for tolerance of European ash (*Fraxinus excelsior*) to dieback disease identified using associative transcriptomics. Sci Rep. 2016;6:19335.

Hebard FV, Islam-Faridi N, Staton ME, Georgi L. Biotechnology of trees: chestnut. In: Tree biotechnology. Boca Raton: CRC Press; 2014. p. 1.

Heffner EL, Sorrells ME, Jannink J-L. Genomic selection for crop improvement. Crop Sci. 2009;49(1):1–12. Crop Science Society of America: 1–12.

Helm D. Natural capital: valuing the planet. New Haven: Yale University Press; 2015.

Howe K, Wood JMD. Using optical mapping data for the improvement of vertebrate genome assemblies. Gigascience. 2015;4:10.

Kafkas S. Whole genome sequencing and high density genetic maps in pistachio reveal a large non-recombining region of sex chromosomes. In: Plant and animal genome XXIV conference. Plant and Animal Genome. 2016. https://pag.confex.com/pag/xxiv/webprogram/Paper21642.html.

Kelly LJ, Leitch AR, Fay MF, Renny-Byfield S, Pellicer J, Macas J, Leitch IJ. Why size really matters when sequencing plant genomes. Plant Ecolog Divers. 2012;5(4):415–25.

Khalturin K, Hemmrich G, Fraune S, Augustin R, Bosch TCG. More than just orphans: are taxonomically-restricted genes important in evolution? Trends Genet. 2009;25(9):404–13.

Kim J, Larkin DM, Asan CQ, Zhang Y, Ge R-L, Auvil L, et al. Reference-assisted chromosome assembly. Proc Natl Acad Sci U S A. 2013;110(5):1785–90.

Kubisiak TL, Nelson CD, Staton ME, Zhebentyayeva T, Smith C, Olukolu BA, Fang G-C, et al. A transcriptome-based genetic map of Chinese chestnut (*Castanea mollissima*) and identification of regions of segmental homology with peach (*Prunus persica*). Tree Genet Genomes. 2013;9(2):557–71.

LaBonte N, Woeste KE. Exploring patterns of sequence variation in regions associated with chestnut blight resistance using whole-genome resequencing of Chinese chestnut (*Castanea mollissima*). In: Plant and animal genome XXIV conference. Plant and Animal Genome. 2016. https://pag.confex.com/pag/xxiv/webprogram/Paper20702.html.

Lesur I, Durand J, Sebastiani F, Gyllenstrand N, Bodénès C, Lascoux M, Kremer A, Vendramin GG, Plomion C. A sample view of the pedunculate oak (*Quercus robur*) genome from the sequencing of hypomethylated and random genomic libraries. Tree Genet Genomes. 2011;7(6):1277–85.

Martínez-García PJ, Crepeau M, Puiu D, Gonzalez-Ibeas D, Stevens K, Whalen J, Butterfield T, et al. The genome sequence of walnut (*Juglans regia* L.) Cv 'Chandler'. In: Plant and animal genome XXIII conference. Plant and Animal Genome. 2015. https://pag.confex.com/pag/xxiii/webprogram/Paper14583.html.

Mathew LS, Spannagl M, Al-Malki A, George B, Torres MF, Al-Dous EK, Al-Azwani EK, et al. A first genetic map of date palm (*Phoenix dactylifera*) reveals long-range genome structure conservation in the palms. BMC Genomics. 2014;15:285.

McKinney LV, Nielsen LR, Hansen JK, Kjær ED. Presence of natural genetic resistance in *Fraxinus excelsior* (Oleraceae) to *Chalara fraxinea* (Ascomycota): an emerging infectious disease. Heredity. 2011;106(5):788–97.

Ming R, Hou S, Feng Y, Yu Q, Dionne-Laporte A, Saw JH, Senin P, et al. The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya* Linnaeus). Nature. 2008;452(7190):991–6.

Naudts K, Chen Y, McGrath MJ, Ryder J, Valade A, Otto J, Luyssaert S. Europe's forest management did not mitigate climate warming. Science. 2016;351(6273):597–600.

Neale DB, Kremer A. Forest tree genomics: growing resources and applications. Nat Rev Genet. 2011;12(2):111–22.

Neale DB, Langley CH, Salzberg SL, Wegrzyn JL. Open access to tree genomes: the path to a better forest. Genome Biol. 2013;14(6):120.

Parra G, Bradnam K, Korf I. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics. 2007;23(9):1061–7.

Pautasso M, Aas G, Queloz V, Holdenrieder O. European ash (*Fraxinus excelsior*) dieback – a conservation biology challenge. Biol Conserv. 2013;158:37–49.

Plomion C, Aury J-M, Amselem J, Alaeitabar T, Barbe V, Belser C, Bergès H, et al. Decoding the oak genome: public release of sequence data, assembly, annotation and publication strategies. Mol Ecol Resour. 2016;16(1):254–65.

Rahman AYA, Usharraj AO, Misra BB, Thottathil GP, Jayasekaran K, Feng Y, Hou S, et al. Draft genome sequence of the rubber tree *Hevea Brasiliensis*. BMC Genomics. 2013;14:75.

Rajaraman S, Salojärvi JT. Silver birch – a model for tree genetics? In: Plant and animal genome XXIII. 2015. https://pag.confex.com/pag/xxiii/webprogram/Paper15896.html.

Resende MDV, Resende Jr MFR, Sansaloni CP, Petroli CD, Missiaggia AA, Aguiar AM, Abad JM, et al. Genomic selection for growth and wood quality in eucalyptus: capturing the missing heritability and accelerating breeding for complex traits in forest trees. New Phytol. 2012;194(1):116–28.

Rowley ER, Fox SA, Bryant DW, Sullivan C, Givan SA, Mehlenbacher SA, Mockler TC. Assembly and characterization of the European Hazelnut (*Corylus avellana* L.) 'Jefferson' transcriptome. Crop Sci. 2012;52:2679–86.

Rowley ER, Genomic resource development for European hazelnut (*Corylus avellana* L.) PhD Thesis, Oregon State University. 2016. http://hdl.handle.net/1957/59368.

Salmon J, Ward SP, Hanley SJ, Leyser O, Karp A. Functional screening of willow alleles in *Arabidopsis* combined with QTL mapping in willow (*Salix*) identifies SxMAX4 as a coppicing responseg. Plant Biotechnol J. 2014;12(4):480–91.

Shen Z, Sun J, Yao J, Wang S, Ding M, Zhang H, Qian Z, et al. High rates of virus-induced gene silencing by tobacco rattle virus in *Populus*. Tree Physiol. 2015;35(9):1016–29.

Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015;31(19):3210–2.

Singh R, Ong-Abdullah M, Low ETL, Manaf MAA, Rosli R, Nookiah R, Ooi LC, et al. Oil palm genome sequence reveals divergence of interfertile species in old and new worlds. Nature. 2013;500(7462):335–9.

Sollars ESA, Harper AL, Kelly LJ, Sambles CM, Ramirez-Gonzalez RH, Swarbreck D, Kaithakottil G, et al. Genome sequence and genetic diversity of European ash trees. Nature. In press, doi:10.1038/nature20786.

Staton M, Best T, Khodwekar S, Owusu S, Xu T, Xu Y, Jennings T, et al. Preliminary genomic characterization of ten hardwood tree species from multiplexed low coverage whole genome sequencing. PLoS One. 2015;10(12):e0145031.

Suarez-Gonzalez A, Hefer CA, Christe C, Corea O, Lexer C, Cronk QCB, Douglas CJ. Genomic and functional approaches reveal a case of adaptive introgression from *Populus balsamifera* (balsam poplar) in *P. trichocarpa* (black cottonwood). Mol Ecol. 2016. doi:10.1111/mec.13539.

VanBuren R, Bryant D, Edger PP, Tang H, Burgess D, Challabathula D, Spittle K, et al. Single-molecule sequencing of the desiccation-tolerant grass *Oropetium thomaeum*. Nature. 2015;527(7579):508–11.

Velasco R, Zharkikh A, Affourtit J, Dhingra A, Cestaro A, Kalyanaraman A, Fontana P, et al. The genome of the domesticated apple (*Malus* x *Domestica* Borkh.). Nat Genet. 2010;42(10):833–9.

Verde I, Abbott AG, Scalabrin S, Jung S, Shu S, Marroni F, Zhebentyayeva T, et al. The high-quality draft genome of peach (*Prunus persica*) identifies unique patterns of genetic diversity, domestication and genome evolution. Nat Genet. 2013;45(5):487–94.

Vezzi F, Narzisi G, Mishra B. Feature-by-feature – evaluating de novo sequence assembly. PLoS One. 2012;7(2):e31002.

Wang N, Thomson M, Bodles WJA, Crawford RMM, Hunt HV, Featherstone AW, Pellicer J, Buggs RJA. Genome sequence of dwarf birch (*Betula nana*) and cross-species RAD markers. Mol Ecol. 2013;22(11):3098–111.

Wegrzyn JL, Lee JM, Tearse BR, Neale DB. TreeGenes: a forest tree genome database. Int J Plant Genomics. 2008;2008:412875.

Wu J, Wang Z, Shi Z, Zhang S, Ming R, Zhu S, Khan MA, et al. The genome of the pear (*Pyrus bretschneideri* Rehd.). Genome Res. 2013;23(2):396–408.

Wu GA, Prochnik S, Jenkins J, Salse J, Hellsten U, Murat F, Perrier X, et al. Sequencing of diverse mandarin, pummelo and orange genomes reveals complex history of admixture during citrus domestication. Nat Biotechnol. 2014;32(7):656–62.

Xu Q, Chen L-L, Ruan X, Chen D, Zhu A, Chen C, Bertrand D, et al. The draft genome of sweet orange (*Citrus sinensis*). Nat Genet. 2013;45(1):59–66.

Zhang J, Li Y, Jia H-X, Li J-B, Huang J, Lu M-Z, Hu J-J. The heat shock factor gene family in *Salix suchowensis*: a genome-wide survey and expression profiling during development and abiotic stresses. Front Plant Sci. 2015;6:748.